Introduction

Why I did chose Kid Creative Dataset?

I have chosen a Multivariate Logistic Regression dataset called “Kid Creative” because,As I mentioned in lab classes, I was looking for prediction of item purchasing in markets to make good business in the future.

What is the dataset about?

This dataset is for logistic regression analysis. Here we got a magazine reseller who wants to sell to his customers. So, with help of information that the reseller got from customers who bought things online before from their market, he made Dataset with 17 features.
THE AIM OF THE MAKING prediction is to decide what magazines to include in **e-mails to customers as a part of an e-mail marketing** campaign.
All of the e-mails that will be sent will go to customers that have previously bought a magazine subscription at MZines4You.com and who have not opted out of receiving e-mails.

How about Dataset information?

Here are the variables that MZines4You.com has on each customer from third-party sources:

Household Income (Income; rounded to the nearest $1,000.00)
Gender (IsFemale = 1 if the person is female, 0 otherwise)
Marital Status (IsMarried = 1 if married, 0 otherwise)
College Educated (HasCollege = 1 if has one or more years of college education, 0 otherwise)
Employed in a Profession (IsProfessional = 1 if employed in a profession, 0 otherwise)
Retired (IsRetired = 1 if retired, 0 otherwise)
Not employed (Unemployed = 1 if not employed, 0 otherwise)
Length of Residency in Current City (ResLength; in years)
Dual Income if Married (Dual = 1 if dual income, 0 otherwise)
Children (Minors = 1 if children under 18 are in the household, 0 otherwise)
Home ownership (Own = 1 if own residence, 0 otherwise)
Resident type (House = 1 if residence is a single family house, 0 otherwise)
Race (White = 1 if race is white, 0 otherwise)
Language (English = 1 is the primary language in the household is English, 0 otherwise)

OUR **TARGET** is BUY column:
Purchased “Kid Creative” (Buy = 1 if purchased “Kid Creative,” 0 otherwise)
Other features are **DATA**.

So the problem of deciding what magazine ads to place in each e-mail boils down to developing an equation for each magazine that predicts the probability that a customer will buy. We are now going to focus on the issue of developing such an equation for one magazine (“Kid Creative”) whose target audience are children between the ages of 9 and 12. In the process of sending out the “experimental” e-mails, the ad for “Kid Creative” was shown in **673 e-mails** to customers and the purchase behavior recorded.

673 e-mails = 673 Instances.
17 Features.

The Data

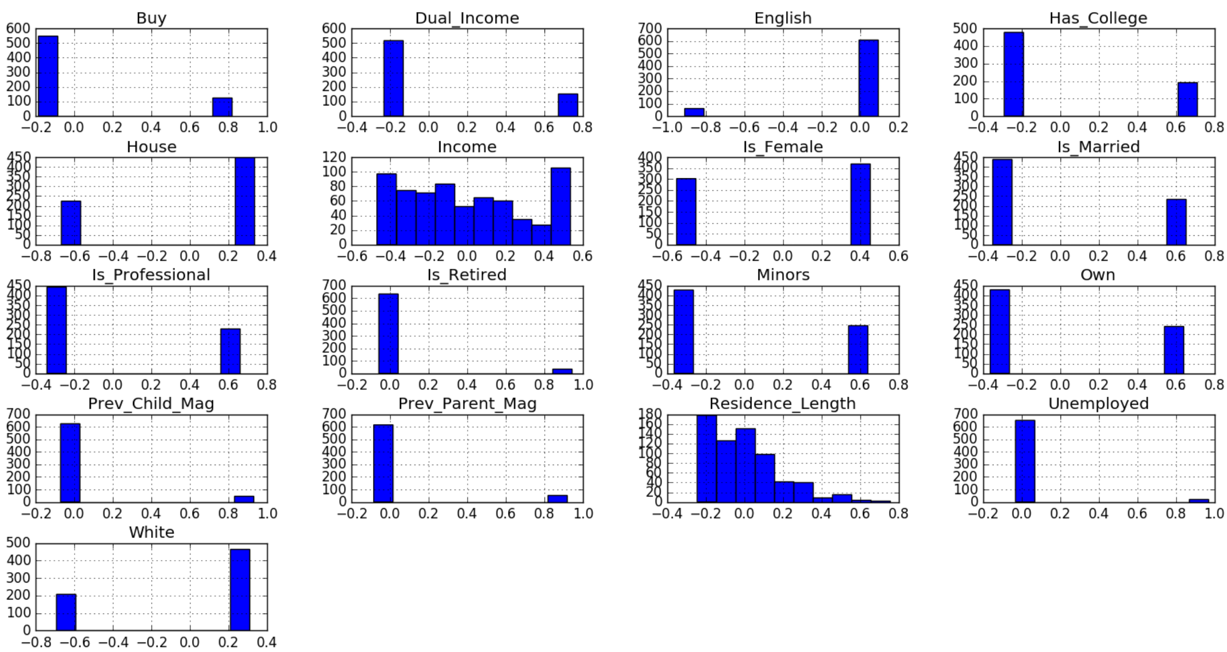
Obs No.	Buy	Income	Is Female	Is Married	Has College	Is Professional	Is Retired	Unemployed	Residence Length	Dual Income	Minors	Own	House	White	English	Prev Child Mag	Prev Parent Mag
1	0	24000	1	0	1	1	0	0	26	0	0	0	1	0	0	0	0
2	1	75000	1	1	1	1	0	0	15	1	0	1	1	1	1	1	0
3	0	46000	1	1	0	0	0	0	36	1	1	1	1	1	1	0	0
4	1	70000	0	1	0	1	0	0	55	0	0	1	1	1	1	1	0
5	0	43000	1	0	0	0	0	0	27	0	0	0	0	1	1	0	1
6	0	24000	1	1	0	0	0	0	41	0	0	1	1	0	0	0	0
7	0	26000	1	1	1	0	1	0	20	0	1	1	1	1	1	0	0
8	0	38000	1	1	0	0	1	0	8	0	0	1	1	1	1	0	0
9	0	39000	1	0	1	1	0	0	17	0	0	0	0	1	1	0	0
10	0	49000	0	1	0	0	1	0	31	0	0	1	1	1	1	0	0
11	1	75000	1	0	1	0	0	0	13	1	0	0	0	1	1	0	1
12	0	31000	1	0	1	0	1	0	51	0	0	0	0	1	1	0	0
13	0	10000	0	0	0	0	0	0	6	0	0	0	0	1	1	0	0
14	0	22000	0	0	0	0	0	0	2	0	0	0	0	1	1	0	0
15	0	39000	0	1	0	0	0	0	24	0	0	1	1	0	0	0	0
16	0	2000	0	1	0	0	0	1	52	1	1	0	1	0	1	0	0
17	1	75000	0	0	0	0	0	0	9	0	0	0	1	1	1	1	0
18	1	69000	1	1	0	0	0	0	0	0	1	1	1	1	1	0	0
19	1	60000	1	1	0	1	0	0	6	1	0	0	0	1	1	1	1
20	0	12000	1	0	0	0	0	0	22	0	0	0	0	1	1	0	0
21	0	42000	1	1	0	0	0	0	46	1	0	1	1	1	1	0	1
22	0	4000	1	0	0	0	0	0	15	0	1	0	1	1	1	0	0
23	0	75000	0	0	0	0	0	0	12	0	1	0	1	1	1	0	0
24	1	45000	1	1	1	0	0	0	20	1	1	1	1	1	1	0	1
25	0	21000	1	0	1	0	0	0	2	0	0	0	0	1	1	0	0
26	1	75000	1	0	1	1	0	0	16	0	0	1	1	0	1	0	0
27	0	23000	0	0	1	1	0	0	2	0	0	0	0	0	1	0	0
28	0	38000	0	0	0	0	1	0	33	0	0	1	1	1	1	0	0
29	0	12000	1	0	0	0	0	0	5	0	0	0	0	1	1	0	0
30	0	46000	1	1	0	0	0	0	16	0	0	1	1	1	1	1	0
31	0	50000	1	1	0	1	0	0	7	1	1	1	1	0	1	0	1

Visualisation of Dataset

Visualisation of dataset is need to see how our features act and how they are placed among the Graph, with their help we can see which ALGORITHMs we should use in this problem.

1.Histogram(all features are in one figure)

Histograms group data into bins and provide you a count of the number of observations in each bin.We can see whet her this graph contains Gaussian and skewed or even has an exponential distribution. Here we see that most of features are categorical, and it's better to use Logit,Knn or DT.

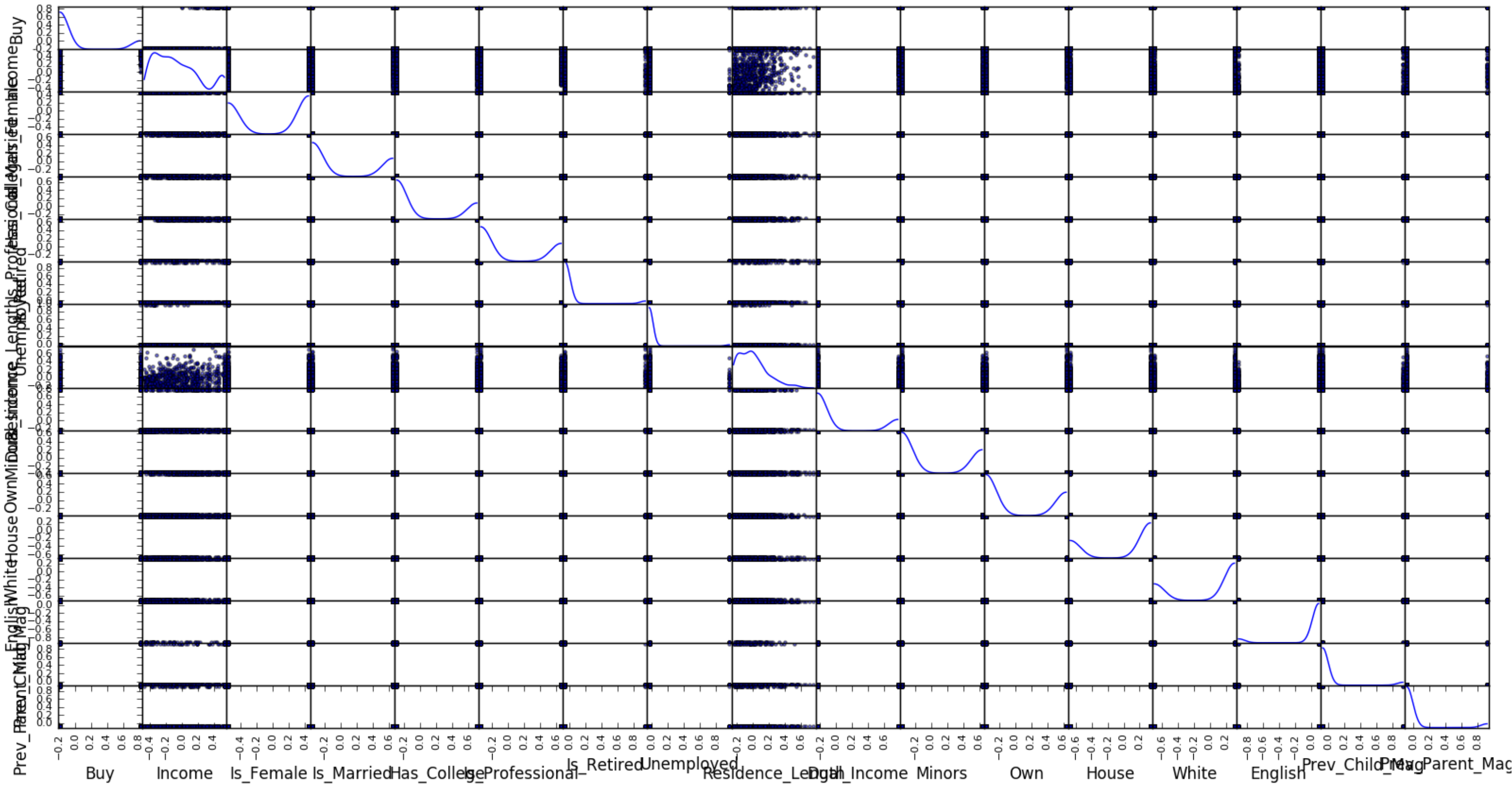


2. Scatter_matrix plot (all features are in one figure)

Correlation gives an indication of how related the changes are between two variables. If two variables change in the same direction they are positively correlated. If the change in opposite directions together (one goes up, one goes down), then they are negatively correlated.

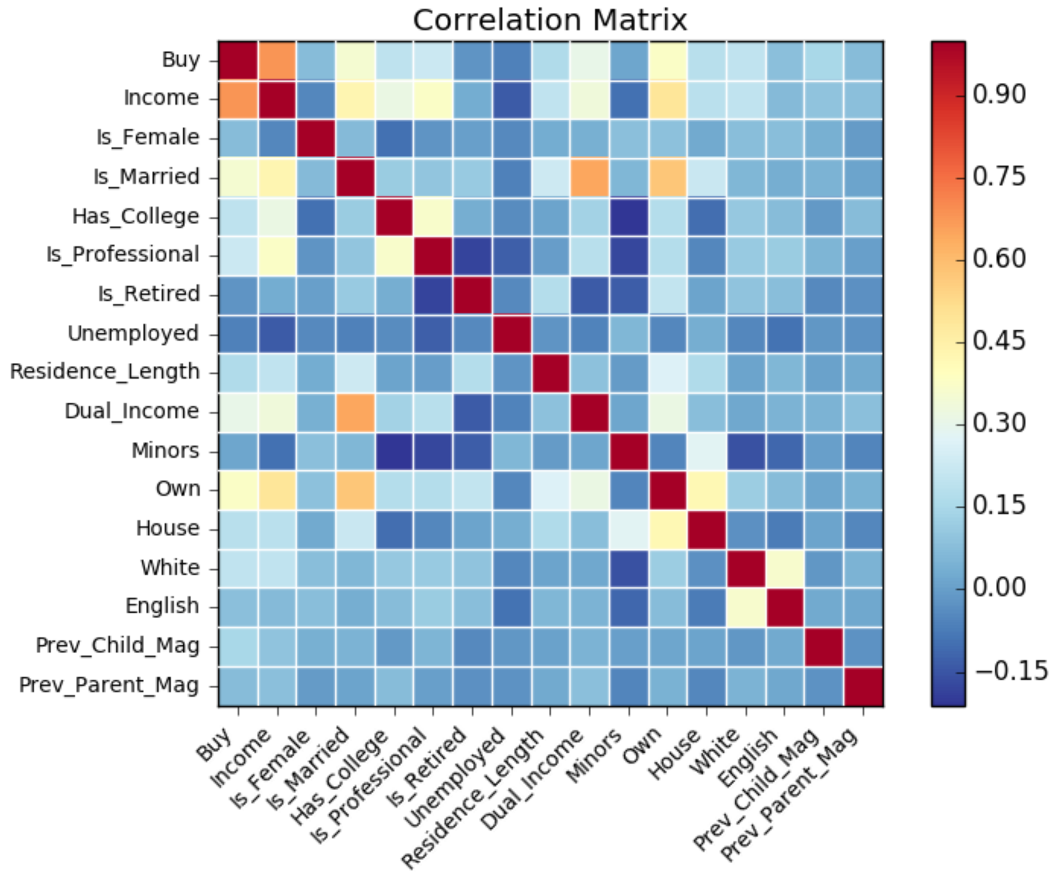
You can calculate the correlation between each pair of attributes. This is called a correlation matrix. You can then plot the correlation matrix and get an idea of which variables have a high correlation with each other.

This is useful to know, because some machine learning algorithms like linear and logistic regression can have poor performance if there are highly correlated input variables in your data.



3. Correlation_matrix (all features are in one figure)

A key thing to remember when working with correlations is never to assume a correlation means that a change in one variable causes a change in another. Buy of magazines and Income of customers have both risen strongly and there is a high correlation between them, but you cannot assume that buying magazines causes people's income (or vice versa).



CODING
Implementation Part

```
1 - Correlations
2 - Visualize correlation figure
3 - Visualize scatter_matrix figure
4 - Visualize only highly correlated features
5 - Visualize histogram figure
6 - Print General accuracy for all appropriate algorithms
7 - Visualize Model implementation
8 - Show newly generated Model's performation and accuracy
9 - Confusion Matrix and for knn and statistics
11 - Get feature Importance using ExtraTreeClassifier
12 - New_Model from Selecting important features, and their accuracy,errors,etc
13 - Get feature Importance using RandomForestClassifier
```

First of All, I did visualisation analysis, and then started generating algorithms, to get the main point, started getting general Accuracies, in order to compare with newly generated ones.

1. General Accuracy

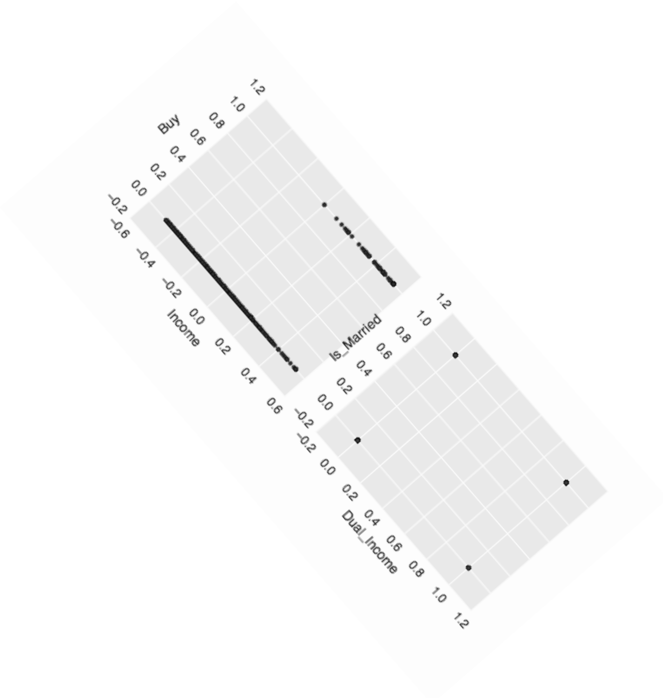
```
accuracy KNN Algorithm: 0.903703703704
accuracy Data Tree: 0.911111111111
accuracy Gaussian Normal: 0.903703703704
accuracy Logistic Regression: 0.866666666667
accuracy SVM : 0.881481481481
accuracy ANN : 0.8
```

As seen in Histogram, actually Gaussian is not suitable for, and Linear regression also, because it's the Classification problem. So, in the next steps, and implementations we give more attention for Logistic Regression.

2. Correlation Analysis.

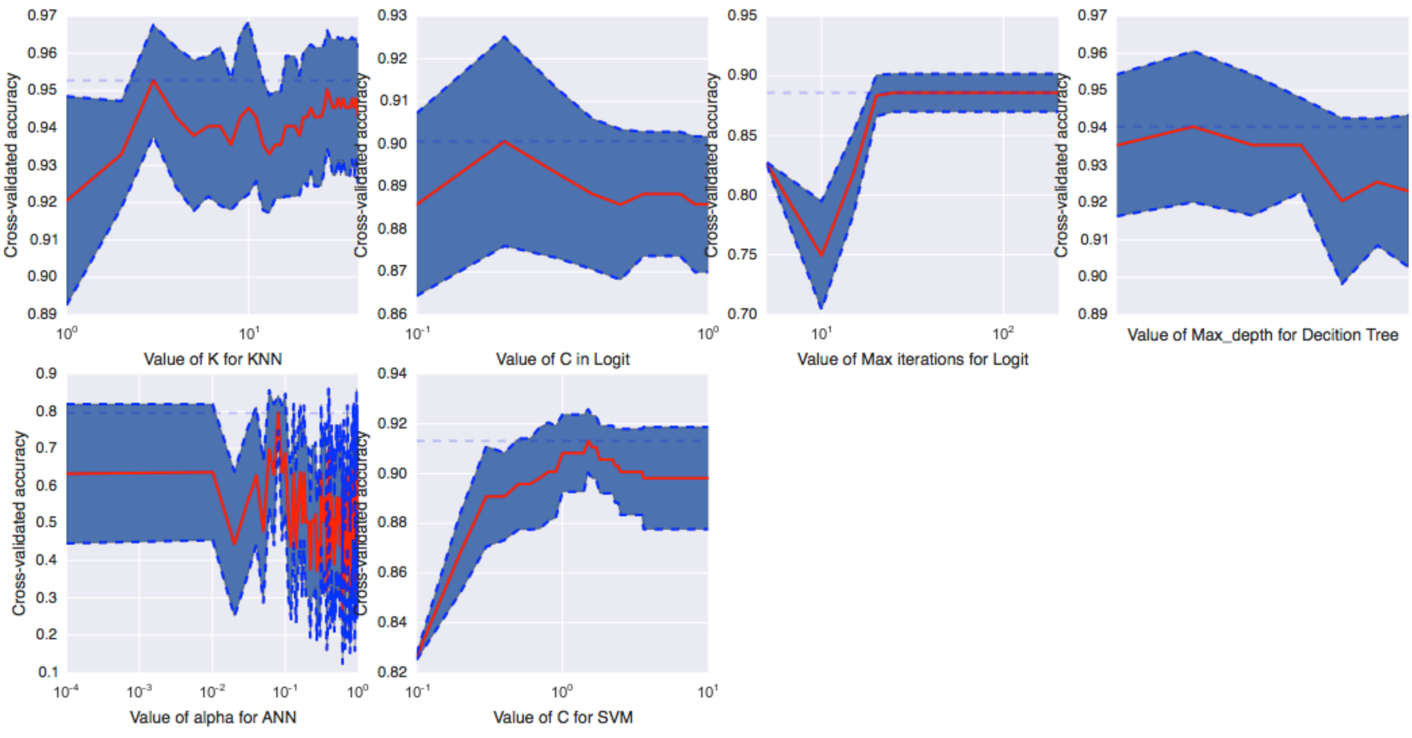
```
corr btw Has_College and Prev_Child_Mag -0.0115544513652
corr btw Is_Professional and Is_Retired -0.178710359915
corr btw Is_Professional and Unemployed -0.129314702316
corr btw Is_Professional and Residence_Length -0.00142683469078
corr btw Is_Professional and Dual_Income 0.183273998205
corr btw Is_Professional and Minors -0.174034886353
corr btw Is_Professional and Own 0.173422820611
corr btw Is_Professional and House -0.0495105899816
corr btw Is_Professional and White 0.113658763264
corr btw Is_Professional and English 0.118363132529
corr btw Is_Professional and Prev_Child_Mag 0.0559439995222
corr btw Is_Retired and Unemployed -0.0445116326433
corr btw Is_Retired and Residence_Length 0.174597592446
corr btw Is_Retired and Dual_Income -0.136240021096
corr btw Is_Retired and Minors -0.134782075497
corr btw Is_Retired and Own 0.209810252656
corr btw Is_Retired and House 0.0132348743772
corr btw Is_Retired and White 0.0964009670672
corr btw Is_Retired and English 0.0783027258789
corr btw Is_Retired and Prev_Child_Mag -0.0440232015724
corr btw Unemployed and Residence_Length -0.0170222381665
corr btw Unemployed and Dual_Income -0.0580786205967
corr btw Unemployed and Minors 0.0595916712872
corr btw Unemployed and Own -0.0464631698157
corr btw Unemployed and House 0.0360827768184
corr btw Unemployed and White -0.0470526983444
corr btw Unemployed and English -0.0921772273959
corr btw Unemployed and Prev_Child_Mag -0.0165291581292
corr btw Residence_Length and Dual_Income 0.0877518114881
corr btw Residence_Length and Minors -0.00739589240091
corr btw Residence_Length and Own 0.270431171807
corr btw Residence_Length and House 0.164684321001
corr btw Residence_Length and White 0.0116357257608
corr btw Residence_Length and English 0.0629311127844
corr btw Residence_Length and Prev_Child_Mag 0.00696015900488
corr btw Dual_Income and Minors 0.0161695700116
corr btw Dual_Income and Own 0.318190217311
corr btw Dual_Income and House 0.0816190757159
corr btw Dual_Income and White 0.0227542052212
corr btw Dual_Income and English 0.0507730102497
corr btw Dual_Income and Prev_Child_Mag 0.0529996294339
corr btw Minors and Own -0.0502730035004
corr btw Minors and House 0.285367452921
corr btw Minors and White -0.158212834879
corr btw Minors and English -0.116099481516
corr btw Minors and Prev_Child_Mag 0.00631157442893
corr btw Own and House 0.421176157806
corr btw Own and White 0.120883587761
corr btw Own and English 0.0766091509623
corr btw Own and Prev_Child_Mag 0.0191833423612
corr btw House and White -0.0266301111946
corr btw House and English -0.0692281707097
corr btw House and Prev_Child_Mag 0.011960726578
corr btw White and English 0.361543795155
corr btw White and Prev_Child_Mag -0.0154662336868
corr btw English and Prev_Child_Mag 0.0271620551882
```

Here we can get features with high correlations abs(x)>0.6, it will help us to deal with feature selection in the next steps.



3. Creating new Models Cross-Validation analysis

Using 10-fold cross-validation analysis, and applying Standard Deviation error we can see and choose the best parameter for new model implementation. So, we select K=27 for KNN, C=0.2 FOR Logit, Max_iter for Logit remains 100, max_depth for DT=4, alpha 0.072 for ANN, and SVM,we chose



4.Accuracy,Error calculation, Validation and Testing

```
**************************************************
Neighbors = 27 is for best model KNeighborsClassifier
C=0.2 is best model for Logistic Regression for
max_depth=4 is best model for DT
Best Feature Selection - SVM 1.5
Best Feature Selection - ANN 0.071
accuracy Of New KNN: 0.903703703704
accuracy Of New LogisticRegression: 0.888888888889
accuracy Of New Decision Tree: 0.911111111111
accuracy Of New SVM: 0.881481481481
accuracy Of New ANN: 0.8

********************LOGISTIC*********************
New Model VS OLD Model For Logit
Logit Variance OLD: 0.87
Logit Variance NEW: 0.89
MSE LOGIT OLD : 0.13
MAE LOGIT OLD : 0.13
RMSE LOGIT OLD : 0.37
MSE LOGIT NEW : 0.11
MAE LOGIT NEW : 0.11
RMSE LOGIT NEW : 0.33

***************************KNN***********************
New Model VS OLD Model For Knn
KNN Variance OLD: 0.90
KNN Variance NEW: 0.90
MSE KNN OLD : 0.10
MAE KNN OLD : 0.10
RMSE KNN OLD : 0.31
MSE KNN NEW : 0.10
MAE KNN NEW : 0.10
RMSE KNN NEW : 0.31
*******************************************************
New Model VS OLD Model For DT
DT Variance OLD: 0.91
DT Variance NEW: 0.91
MSE DT OLD : 0.09
MAE DT OLD : 0.09
RMSE DT OLD : 0.30
MSE DT NEW : 0.09
MAE DT NEW : 0.09
RMSE DT NEW : 0.30
*******************************************************
New Model VS OLD Model For SVM
SVM Variance OLD: 0.88
SVM Variance NEW: 0.88
MSE SVM OLD : 0.09
MAE SVM OLD : 0.09
RMSE SVM OLD : 0.30
MSE SVM NEW : 0.09
MAE SVM NEW : 0.09
RMSE SVM NEW : 0.30
*******************************************************
```

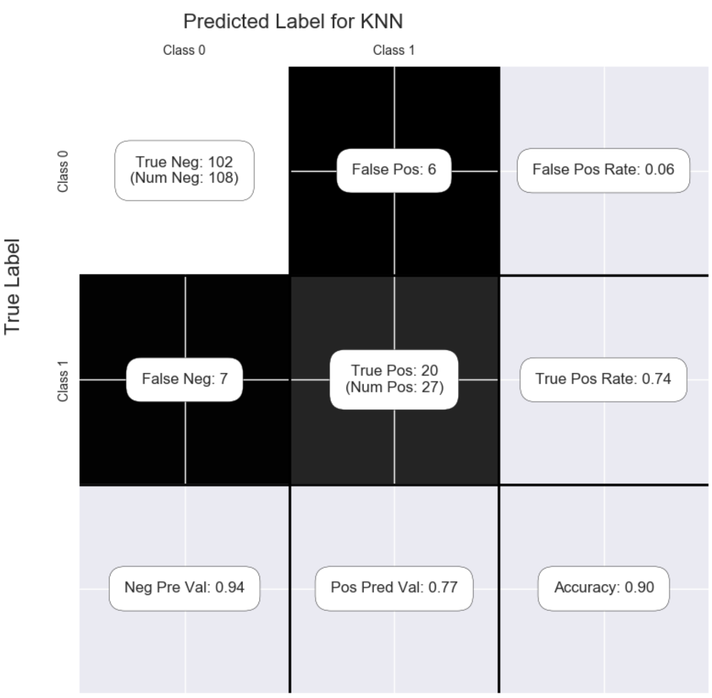
After Implementing and Testing new models, We

1.Logistic Accuracy score increased from 0.866 to 0.8888 The MSE and RMSE error decreased from 0.13->-0.11 and 0.37->0.33, The Strength of accuracy(Variance) increased from 0.87 to 0.89.
2.KNN's accuracy and error_metrics remains same
3.SVM's Accuracy and error_metrics remains same
4.ANN' Accuracy from 0.22->0.79 and Variance from 0.22->0.8
5.DT remains same

```
****************************************
New Model VS OLD Model For ANN
ANN Variance OLD: 0.22
ANN Variance NEW: 0.79
MSE ANN OLD : 0.09
MAE ANN OLD : 0.09
RMSE ANN OLD : 0.30
MSE ANN NEW : 0.09
MAE ANN NEW : 0.09
RMSE ANN NEW : 0.30
********************TEST best parameters
accuracy knn TEST: 0.896296296296
accuracy logistic TEST: 0.866666666667
accuracy SVM TEST: 0.866666666667
accuracy DT TEST: 0.896296296296
accuracy ANN TEST: 0.777777777778
```

5.Confusion Matrix for different Algorithms

KNN and Logit Algorithm.Confusion Matrix: It is nothing but a tabular representation of Actual vs Predicted values.
This helps us to find the accuracy of the model and avoid overfitting. This is how it looks like:'''

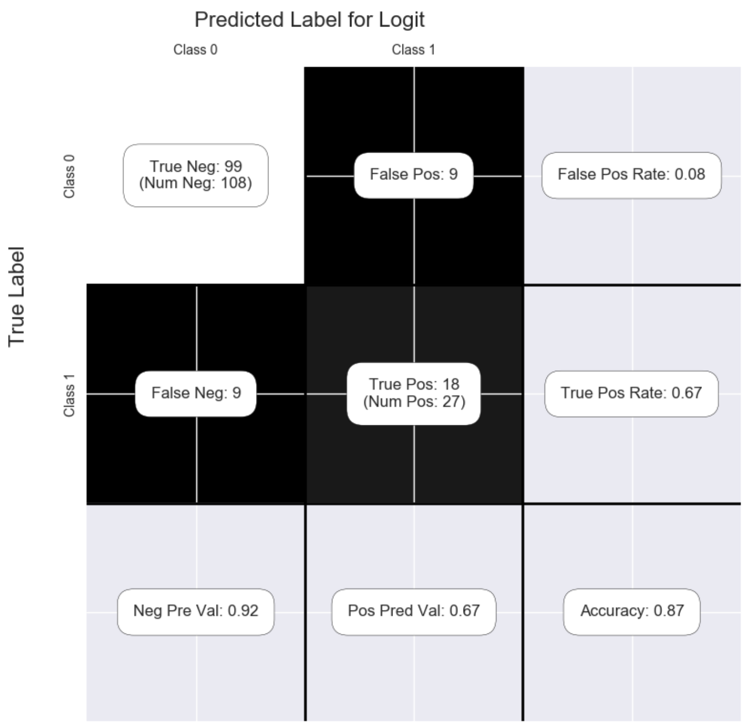



```
Class Statistics:

Classes                                         0          1
Population                                    135        135
P: Condition positive                         108         27
N: Condition negative                          27        108
Test outcome positive                         109         26
Test outcome negative                          26        109
TP: True Positive                             102         20
TN: True Negative                              20        102
FP: False Positive                              7          6
FN: False Negative                              6          7
TPR: (Sensitivity, hit rate, recall)     0.944444   0.740741
TNR=SPC: (Specificity)                   0.740741   0.944444
PPV: Pos Pred Value (Precision)           0.93578   0.769231
NPV: Neg Pred Value                      0.769231    0.93578
FPR: False-out                           0.259259  0.0555556
FDR: False Discovery Rate               0.0642202   0.230769
FNR: Miss Rate                          0.0555556   0.259259
ACC: Accuracy                            0.903704   0.903704
F1 score                                 0.940092   0.754717
MCC: Matthews correlation coefficient    0.695027   0.695027
Informedness                             0.685185   0.685185
Markedness                               0.705011   0.705011
Prevalence                                    0.8        0.2
LR+: Positive likelihood ratio            3.64286    13.3333
LR-: Negative likelihood ratio              0.075    0.27451
DOR: Diagnostic odds ratio                48.5714    48.5714
FOR: False omission rate                 0.230769  0.0642202
True positives: 20
True negatives: 102
False negatives: 7
False positives 6
             precision    recall  f1-score   support

          0       0.94      0.94      0.94       108
          1       0.77      0.74      0.75        27

avg / total       0.90      0.90      0.90       135
```




```
Class Statistics:

Classes                                       0          1
Population                                  135        135
P: Condition positive                       108         27
N: Condition negative                        27        108
Test outcome positive                       108         27
Test outcome negative                        27        108
TP: True Positive                            99         18
TN: True Negative                            18         99
FP: False Positive                            9          9
FN: False Negative                            9          9
TPR: (Sensitivity, hit rate, recall)   0.916667   0.666667
TNR=SPC: (Specificity)                 0.666667   0.916667
PPV: Pos Pred Value (Precision)        0.916667   0.666667
NPV: Neg Pred Value                    0.666667   0.916667
FPR: False-out                         0.333333  0.0833333
FDR: False Discovery Rate             0.0833333   0.333333
FNR: Miss Rate                        0.0833333   0.333333
ACC: Accuracy                          0.866667   0.866667
F1 score                               0.916667   0.666667
MCC: Matthews correlation coefficient  0.583333   0.583333
Informedness                           0.583333   0.583333
Markedness                             0.583333   0.583333
Prevalence                                  0.8        0.2
LR+: Positive likelihood ratio             2.75          8
LR-: Negative likelihood ratio            0.125   0.363636
DOR: Diagnostic odds ratio                   22         22
FOR: False omission rate               0.333333  0.0833333
True positives: 18
True negatives: 99
False negatives: 9
False positives 9
             precision    recall  f1-score   support

          0       0.92      0.92      0.92       108
          1       0.67      0.67      0.67        27

avg / total       0.87      0.87      0.87       135
```

6.Future Selection and Implementation

Feature selection

Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.

Improves Accuracy: Less misleading data means modeling accuracy improves.

Reduces Training Time: Less data means that algorithms train faster.'''

ExtraTreeClassifier:

Extra Trees model to the data

: display the relative importance of each attribute

: suggests that N features are informative

In our case N is 2, here we get only 2 features which are informative

Those: Income, Residence_Length features

```
11 - Get feature Importance using ExtraTreeClassifier
12 - New_Model from Selecting important features, and their accuracy,error
c
13 - Get feature Importance using RandomForestClassifier
x - To exit
Enter The command: 11
[ 0.49874222  0.02433373  0.05235889  0.04199756  0.03057932  0.00832098
  0.00579305  0.09479574  0.02138358  0.02583343  0.070546    0.02644579
  0.04707225  0.00714538  0.02151158  0.02314053]
1 - Correlations
```

A random forest regressor:

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

#Using RandomForestRegressor, we will select features with high values, which mean those features are important
#but, it takes huge amount of Time to implement 16 features(17th is Target).
Univariate feature selection is in general best to get a better understanding of the data, its structure and characteristics. It can work for selecting top features for model improvement in some settings, but since it is unable to remove redundancy (for example selecting only the best feature among a subset of strongly correlated features)

```
13 - Get feature Importance using RandomForestClassifier
x - To exit
Enter The command: 13
You have to wait until it performes... about 3-5minutes...
[(0.477, 'Income'), (-0.025, 'Own'), (-0.025, 'Is_Married'), (-0.043, 'Dual_Income'), (-0.073, 'Is_Professional'), (-0.096, 'House'), (-0.099, 'Has_College'), (-0.109, 'White'), (-0
.115, 'Unemployed'), (-0.117, 'Is_Female'), (-0.128, 'Is_Retired'), (-0.129, 'Prev_Child_Mag'), (-0.133, 'English'), (-0.137, 'Residence_Length'), (-0.138, 'Prev_Parent_Mag'), (-0.1
41, 'Minors')]
```

New_Model_Implementation(Command 12)

After Selecting 2 Important features "INCOME" and "Residence_Length", Our model performed very well, 1st: It run very fast, 2nd: Gave us best prediction with less error. 3rd: More accurate
Let's see each algorithm one by one:

<u>All algorithms with best parameters(Cross-validated)</u>

1. Logistic Regression:
   MSE decreased to 0.07 from 0.11(previous best model)
   RMSE decreased to 0.26 from 0.33(previous best model)
   Accuracy score is 0.93, previous was 0.8888, 0.042 more.
2. KNN:
   MSE decreased to 0.06 from 0.10(previous best model)
   RMSE decreased to 0.24 from 0.31(previous best model)
   Accuracy score is 0.94, previous was 0.903, 0.04 more
3. DT:
   MSE decreased to 0.07 from 0.09(previous best model)
   RMSE decreased to 0.26 from 0.30(previous best model)
   Accuracy score is 0.93, previous was 0.91.
4. SVM:
   MSE decreased to 0.07 from 0.01(previous best model)
   RMSE decreased to 0.26 from 0.33(previous best model)
   Accuracy score is 0.93, previous was 0.88, 0.042 more.
5. ANN:
   MSE increased to 0.15 from 0.09(previous best model)
   RMSE increased to 0.39 from 0.30(previous best model)
   Accuracy score is INCREASED to **0.851**, previous was 0.77, 0.042 more.

Image provided below

```
12 - New_Model from Selecting important features, and their accuracy,errors,etc
13 - Get feature Importance using RandomForestClassifier
x - To exit
Enter The command: 12
Best Feature Selection - Logistic Regression
accuracy Of New LogisticRegression: 0.930693069307
MSE3: 0.07
MAE3: 0.07
RMSE3: 0.26

Best Feature Selection - Decision Tree
accuracy Of New Decision Tree: 0.930693069307
MSE3: 0.07
MAE3: 0.07
RMSE3: 0.26

Best Feature Selection - KNN
accuracy Of New KNN: 0.940594059406
MSE3: 0.06
MAE3: 0.06
RMSE3: 0.24
Best Feature Selection - SVM
accuracy Of New SVM: 0.930693069307
MSE3: 0.07
MAE3: 0.07
RMSE3: 0.26
Best Feature Selection - ANN
accuracy Of New ANN: 0.851485148515
MSE3: 0.15
MAE3: 0.15
RMSE3: 0.39
```