Kennesaw State University
IS 8935 Business Intelligence: Traditional & Big Data Analytics
Dr. Reza Vaezi
Assignment 7
March 4, 2019

Logistic Regression Analysis for Active Life Style

By Karis Kim

Executive Summary

The goal of this report is to provide insight for the NGO to start a targeted campaign to help individuals adopt an active life style. Based on confidence ratings of logistic regression analysis, the report identifies 25 individuals most likely to adopt an active life style. Those 25 individuals are under 30 single females who eat breakfast bars for breakfast. The report recommends starting the campaign with these 25 and then extending the campaign to other individuals who fall in the category, as well as to consider more attributes like smoking, drinking and exercise to run further/more informed analysis.

CONTENTS

BUSINESS UNDERSTANDING

An NGO that promotes an active life style is starting a targeted campaign to help individuals adopt an active life style. The goal of this report is to identify a list of individuals who are likely not following an active life style and at the same time are the most likely to adopt an active life style if contacted. With the findings of the prediction in this report, these inactive individuals will be directly contacted with a campaign for active life style to promote health and well-being.

DATA UNDERSTANDING

The ActiveStatus data sets contain various attributes that may potentially impact whether individuals pursue an active life style or not. The target/label attribute is Active, which is the dependent variable to predict whether or not an individual is likely to adopt an active life style.

ATTRIBUTE INFORMATION

	Attribute	Description
1	ID	ID
2	agecat	Age Category (1= under30, 2= 31-45, 3= 46-60, 4= over 60)
3	age	Age
4	gender	0= male, 1= female
5	marital	Marital status (0 Unmarried, 1 Married)
6	bfast	Breakfast (1= breakfast bar, 2= oatmeal, 3= cereal)
7	income	hourly income in dollars
8	active	active life style (0= no, 1= yes) - Dependent variable

The data set is divided into ActiveStatus_Training data set and ActiveStatus_Scoring data set for logistic regression analysis. The ActiveStatus_Training data set contains 198 examples and 8 attributes, while the ActiveStatus_Scoring data set contains 682 examples and all attributes except the Active attribute, which will be the label or predictor attribute.

DATA ASSUMPTIONS

Data set only contains 3 possible answers for the breakfast attribute and works under the assumption that no other types of breakfast foods were/are consumed by the subjects of the data set to indicate active life style.

It is assumed that age categories were accurately assigned respective to the individual actual ages.

Figure 1. Descriptive statistics of Training data set

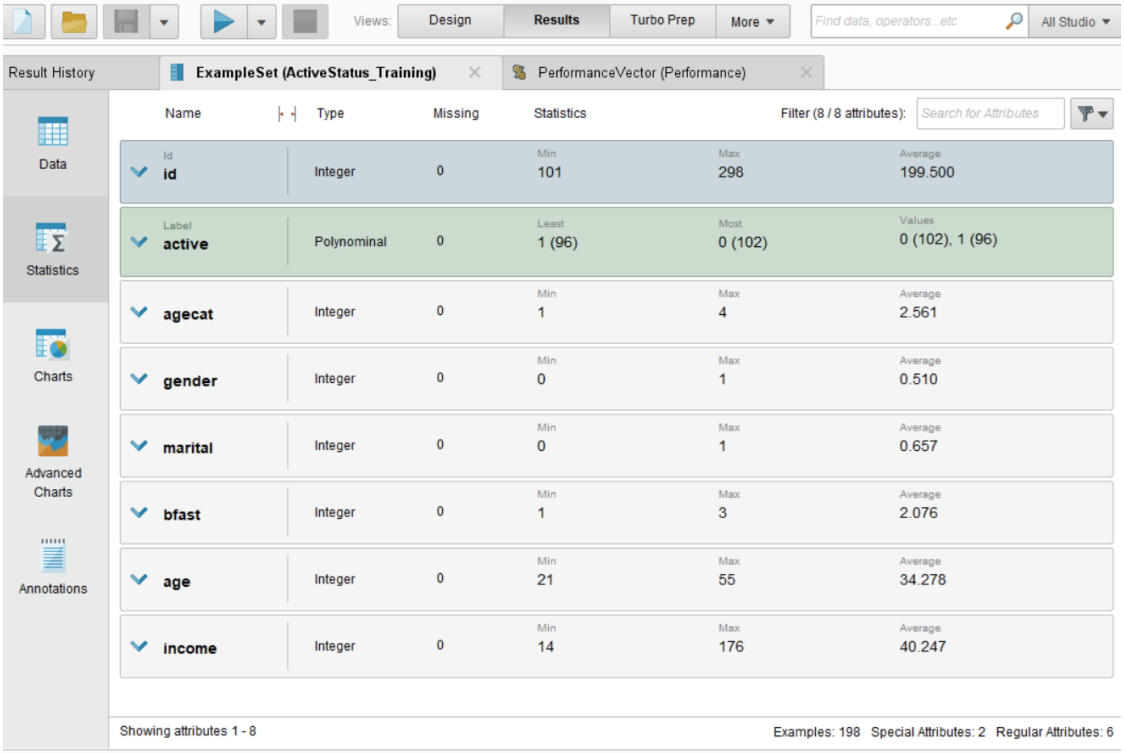
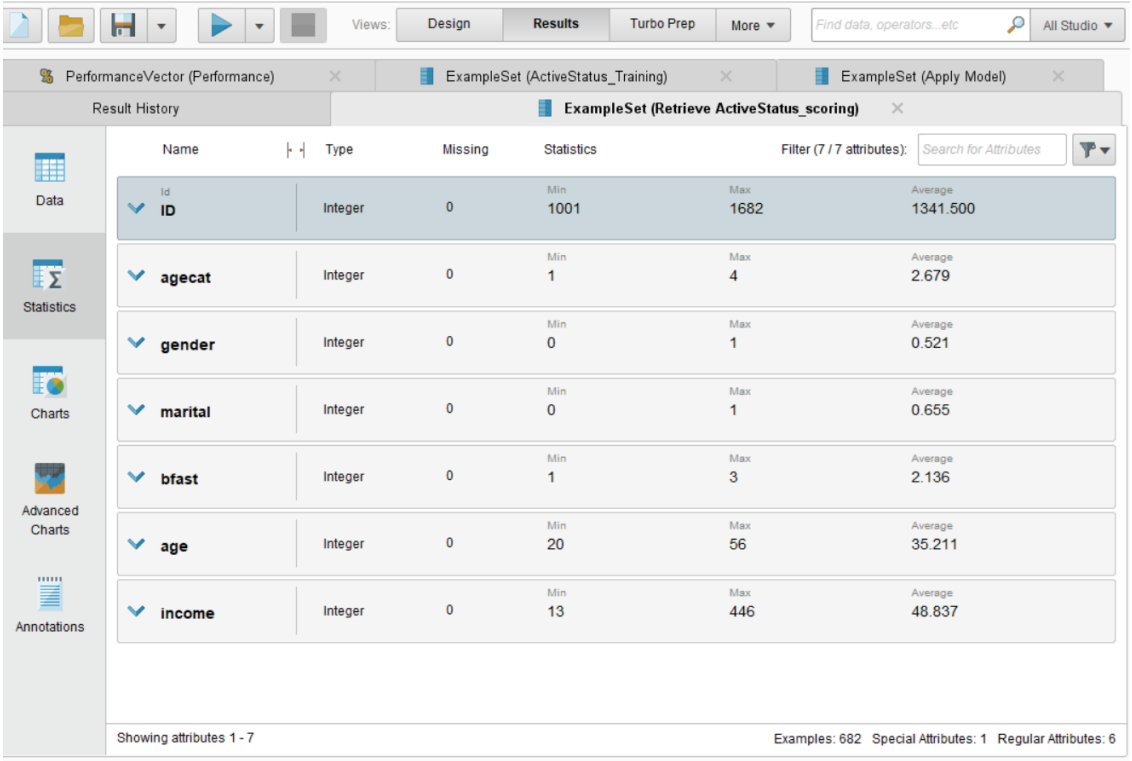


Figure 2. Descriptive statistics of Scoring data set



DATA PREPARATION

Data Type Transformation: At data import of training set, the *ID* attribute was designated as ID and the *active* attribute was designated as label. No transformation to data types was needed. At data import of the scoring set, the ID attribute was designated as ID, and no transformation of data types was needed.
Data Preparation of Missing Values: No missing values were found in the data set.

DATA RANGES FOR ATTRIBUTES IN SCORING SET

All data ranges for attributes in the scoring set must be within the range of those in the corresponding training set. Comparison of the scoring and training data sets showed that for attribute *age*, the training data set had a range of 21 to 55, while the scoring data set had 20 to 56. Similarly, the attribute *income* in training data set has a range of 14 to 176, but the scoring data set has 13 to 446.
Data Preparation for Out of Range Attributes: In RapidMiner, Filter Examples operator was used to remove those out of range observations from the data set. In Filter Examples Parameters, click Add Filters, and in the Create Filters window, enter the range of values of *age* and *income* from the training data set, so that all examples within those values may be passed through to the output (see figure 3, 4).

Figure 3. Filter Examples operator to remove observations out of range from scoring set

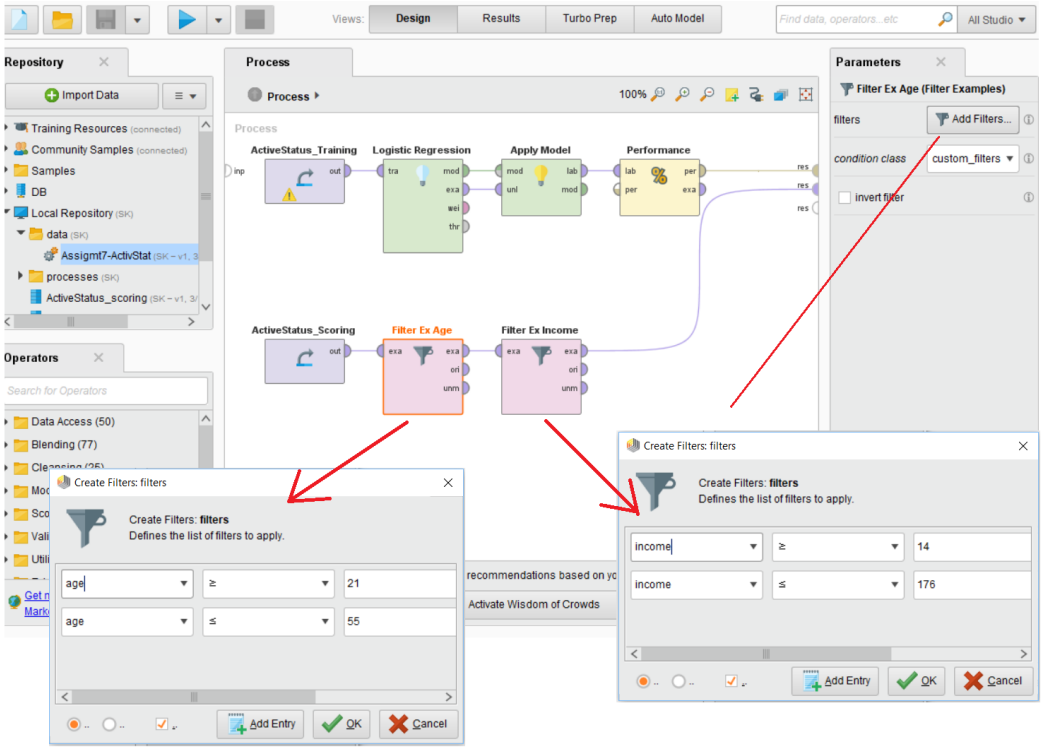
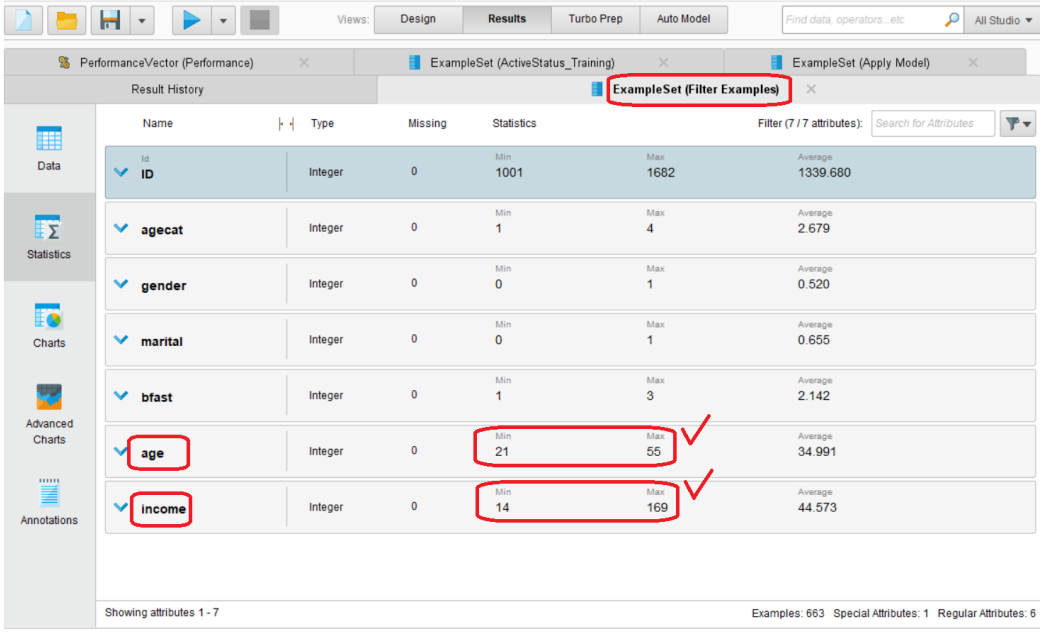


Figure 4. Results showing *age* and *income* examples in scoring set after removing out of range observations



MODELING

Following data preparation, the logistic regression modeling process will involve running the Logistic Regression operator on the training set, then running the Apply Model operator, and the Performance (Classification) operator to measure the performance of the logistic regression model. Then the model can be deployed on the scoring set.

LOGISTIC REGRESSION

First, the Logistic Regression operator is added to the training set with parameters at default setting (see figure 5).

APPLY MODEL ON THE TRAINING SET

Next, the Apply Model operator is added after the Logistic Regression operator to apply the model to the training set (see figure 5).

Then the process can be run to generate the following results in Figure 6. In the Results view, in the Logistic Regression tab, doubling clicking the columns will sort values ascending or descending order.

PERFORMANCE (CLASSIFICATION)

Before applying the Logistic Regression model to the scoring set, the Performance (Classification) operator is added to measure how accurate the model is (see figure 5). The overall accuracy in this logistic regression model is 64.65% as shown in Figure 7.

Figure 5. RapidMiner process for logistic regression and performance

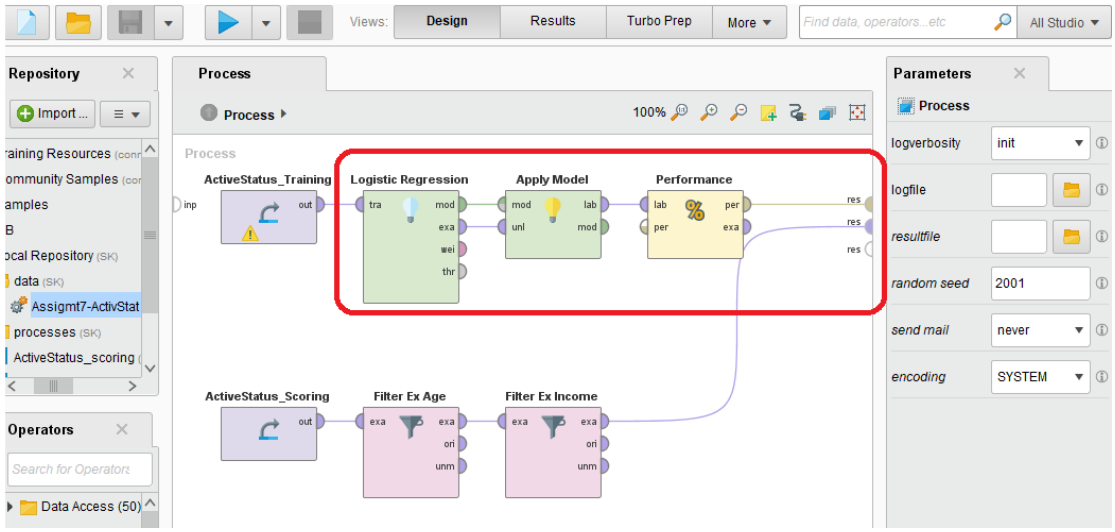
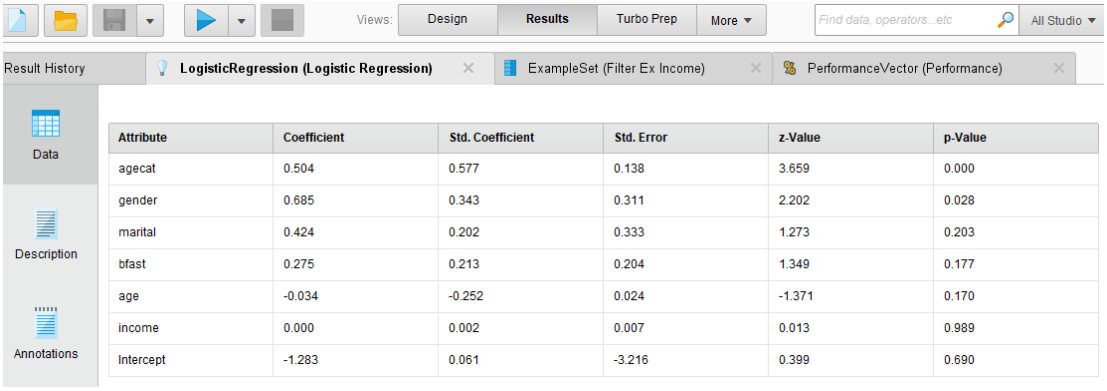
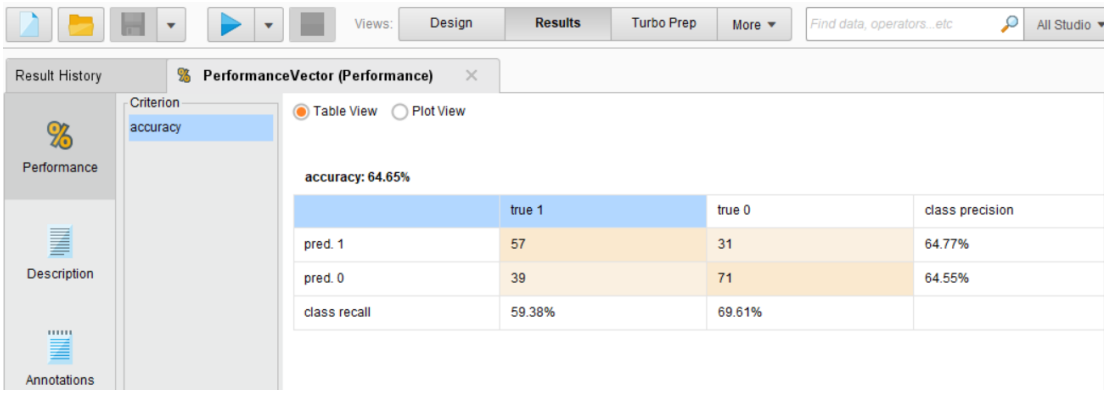


Figure 6. Results of logistic regression on training set



Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
agecat	0.504	0.577	0.138	3.659	0.000
gender	0.685	0.343	0.311	2.202	0.028
marital	0.424	0.202	0.333	1.273	0.203
bfast	0.275	0.213	0.204	1.349	0.177
age	-0.034	-0.252	0.024	-1.371	0.170
income	0.000	0.002	0.007	0.013	0.989
Intercept	-1.283	0.061	-3.216	0.399	0.690

Figure 7. RapidMiner Result of Performance of Logistic Regression on training set



accuracy: 64.65%

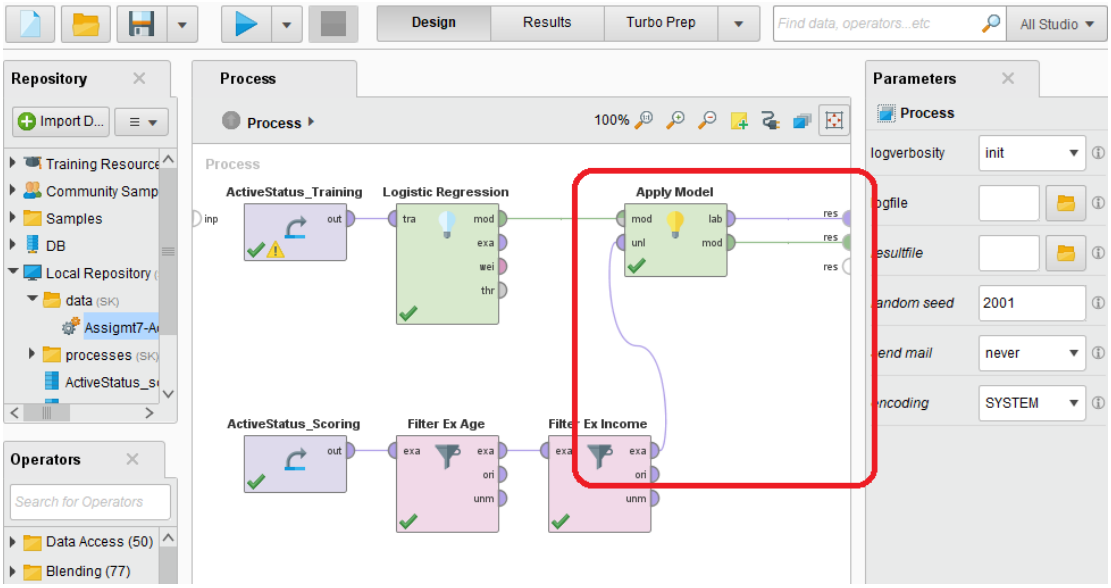
	true 1	true 0	class precision
pred. 1	57	31	64.77%
pred. 0	39	71	64.55%
class recall	59.38%	69.61%	

APPLY MODEL ON THE SCORING SET

Now the model will be applied to the scoring set, by connecting the Apply Model operator's unlabeled input port with the examples from the scoring data set stream (see figure 8). The process of applying the Logistic Regression

model on the scoring set will yield a data table of the example set with the prediction(active) column containing prediction active(1) or inactive(0) per ID for the 663 examples in the scoring set (see figure 9).

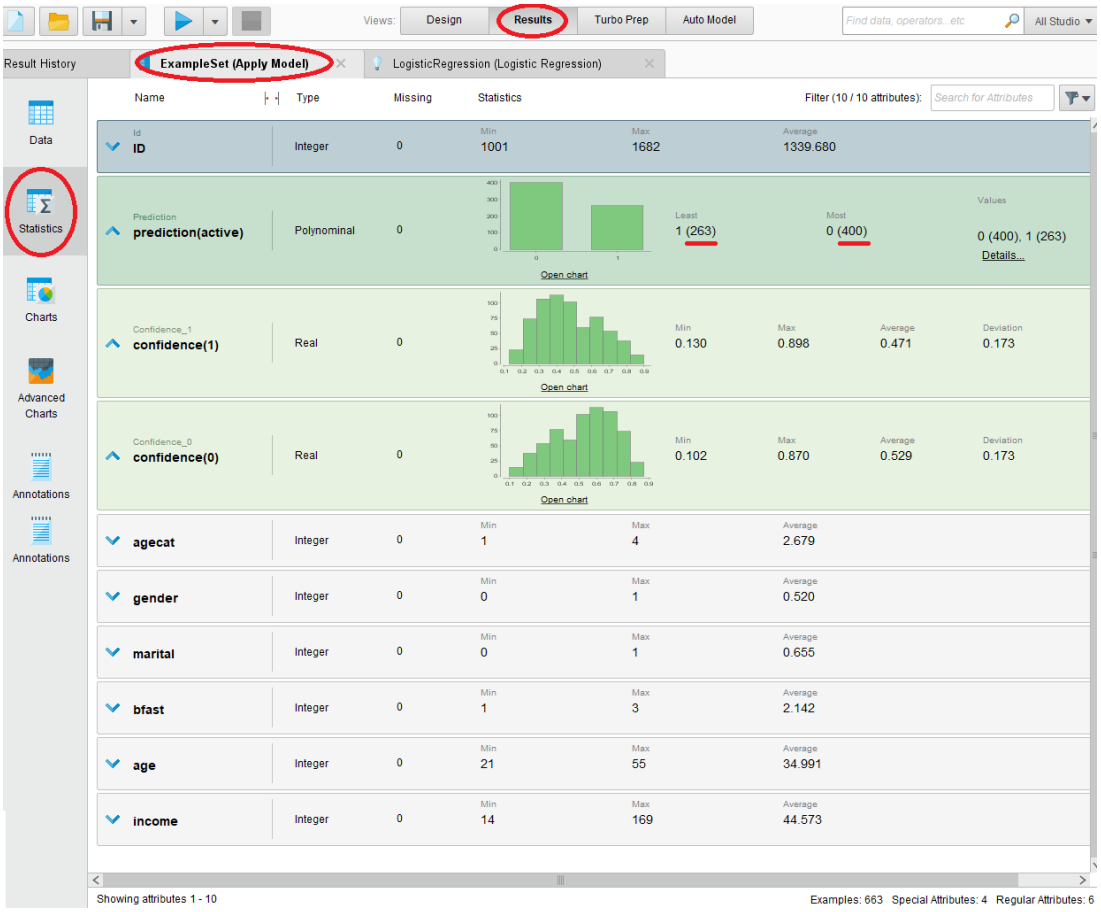
Figure 8. Connect Apply Model operator to examples from the scoring set



RESULTS

The following Figure 9 shows the results of the logistic regression model applied on the scoring set, with the prediction(active) highlighted in the green column. The logistic regression model predicted that 263 would be active and 400 would be inactive.

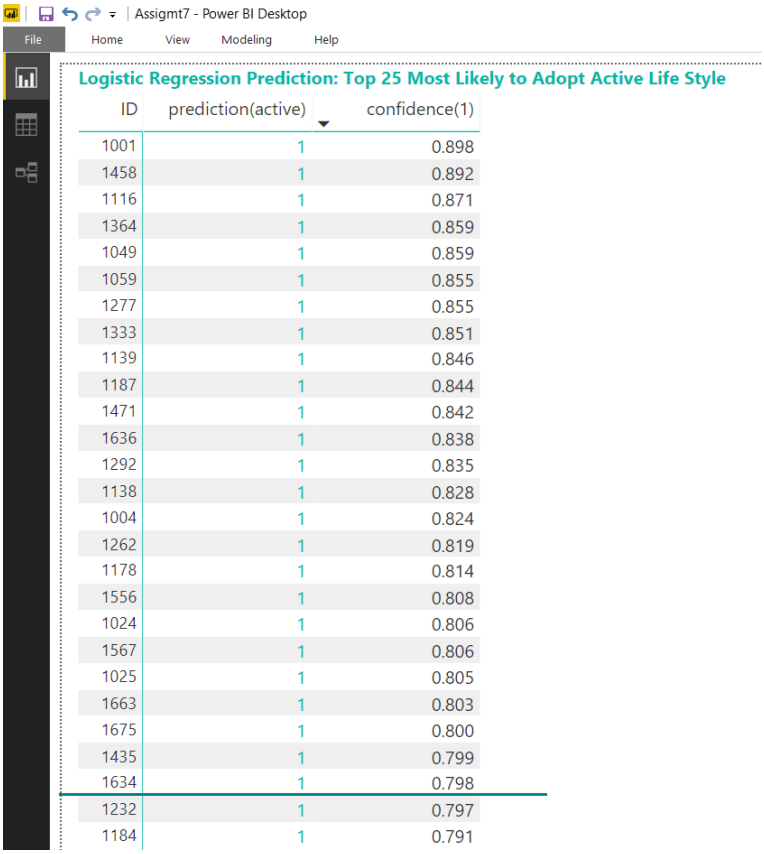
Figure 9. Results showing descriptive statistics on scoring set with prediction(active)



EVALUATION OF FINDINGS

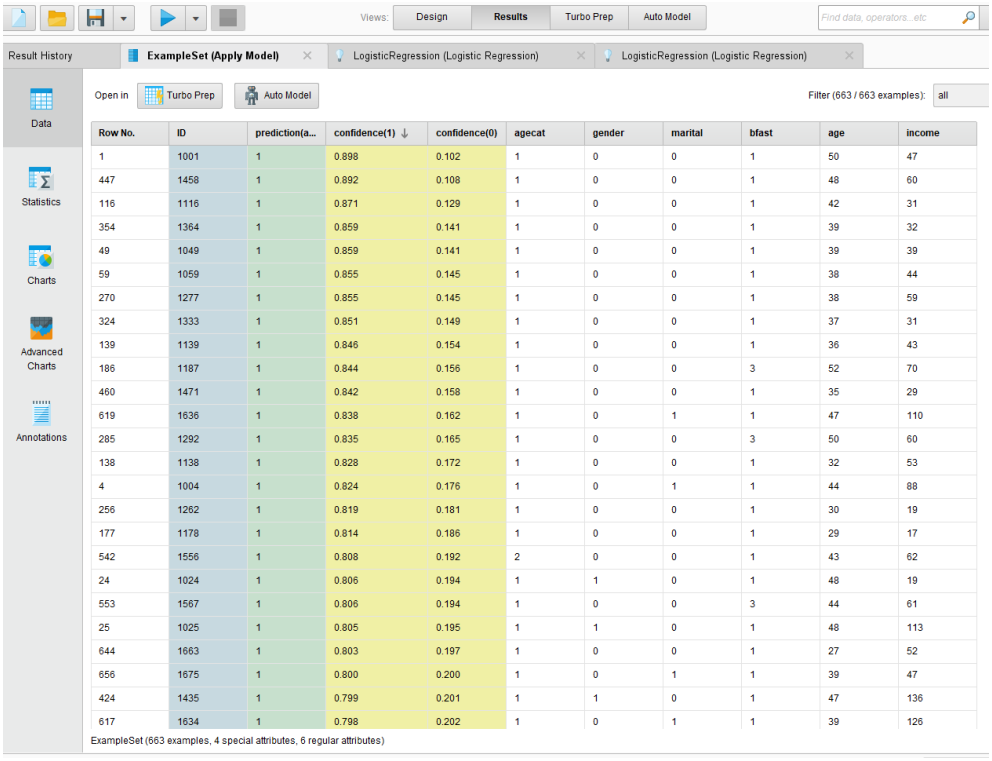
- The following Figure 10 and 11 shows a list of the top 25 individuals predicted to be most likely adopt an active life style. This list was sorted based on the confidence rating for active life style.
- Figure 12 shows that all but 1 of these 25 individuals is in the under 30 age category (agecat =1).
- Figure 13 shows that all but 3 of these 25 individuals are female.
- Figure 14 shows that all but 4 of these 25 individuals are single.
- Figure 15 shows that all but 3 of these 25 individuals eat breakfast bars.

Figure 10. Power BI table of top 25 predicted to be most likely to adopt an active life style



ID	prediction(active)	confidence(1)
1001	1	0.898
1458	1	0.892
1116	1	0.871
1364	1	0.859
1049	1	0.859
1059	1	0.855
1277	1	0.855
1333	1	0.851
1139	1	0.846
1187	1	0.844
1471	1	0.842
1636	1	0.838
1292	1	0.835
1138	1	0.828
1004	1	0.824
1262	1	0.819
1178	1	0.814
1556	1	0.808
1024	1	0.806
1567	1	0.806
1025	1	0.805
1663	1	0.803
1675	1	0.800
1435	1	0.799
1634	1	0.798
1232	1	0.797
1184	1	0.791

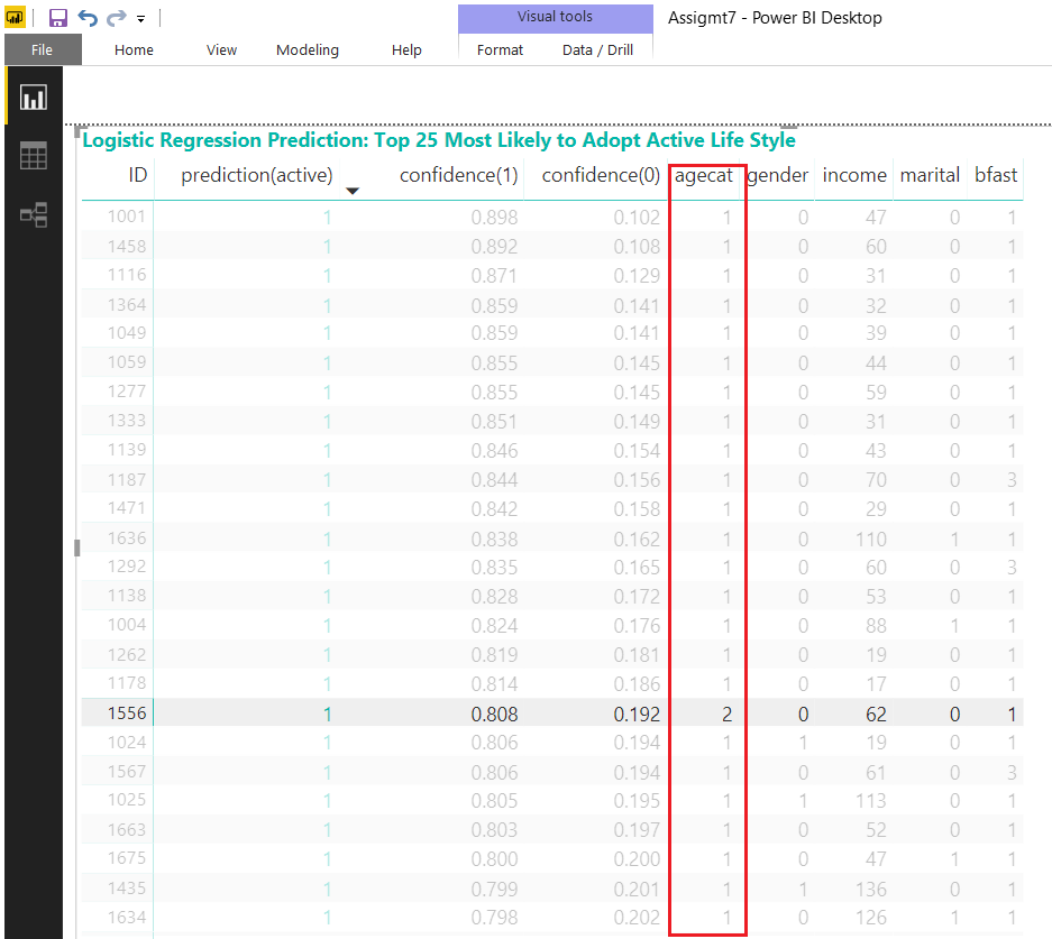
Figure 11. RapidMiner table of top 25 predicted to be most likely to adopt an active life style



Row No.	ID	prediction(a...	confidence(1)	confidence(0)	agecat	gender	marital	bfast	age	income
1	1001	1	0.898	0.102	1	0	0	1	50	47
447	1458	1	0.892	0.108	1	0	0	1	48	60
116	1116	1	0.871	0.129	1	0	0	1	42	31
354	1364	1	0.859	0.141	1	0	0	1	39	32
49	1049	1	0.859	0.141	1	0	0	1	39	39
59	1059	1	0.855	0.145	1	0	0	1	38	44
270	1277	1	0.855	0.145	1	0	0	1	38	59
324	1333	1	0.851	0.149	1	0	0	1	37	31
139	1139	1	0.846	0.154	1	0	0	1	36	43
186	1187	1	0.844	0.156	1	0	0	3	52	70
460	1471	1	0.842	0.158	1	0	0	1	35	29
619	1636	1	0.838	0.162	1	0	1	1	47	110
285	1292	1	0.835	0.165	1	0	0	3	50	60
138	1138	1	0.828	0.172	1	0	0	1	32	53
4	1004	1	0.824	0.176	1	0	1	1	44	88
256	1262	1	0.819	0.181	1	0	0	1	30	19
177	1178	1	0.814	0.186	1	0	0	1	29	17
542	1556	1	0.808	0.192	2	0	0	1	43	62
24	1024	1	0.806	0.194	1	1	0	1	48	19
553	1567	1	0.806	0.194	1	0	0	3	44	61
25	1025	1	0.805	0.195	1	1	0	1	48	113
644	1663	1	0.803	0.197	1	0	0	1	27	52
656	1675	1	0.800	0.200	1	0	1	1	39	47
424	1435	1	0.799	0.201	1	1	0	1	47	136
617	1634	1	0.798	0.202	1	0	1	1	39	126

ExampleSet (663 examples, 4 special attributes, 6 regular attributes)

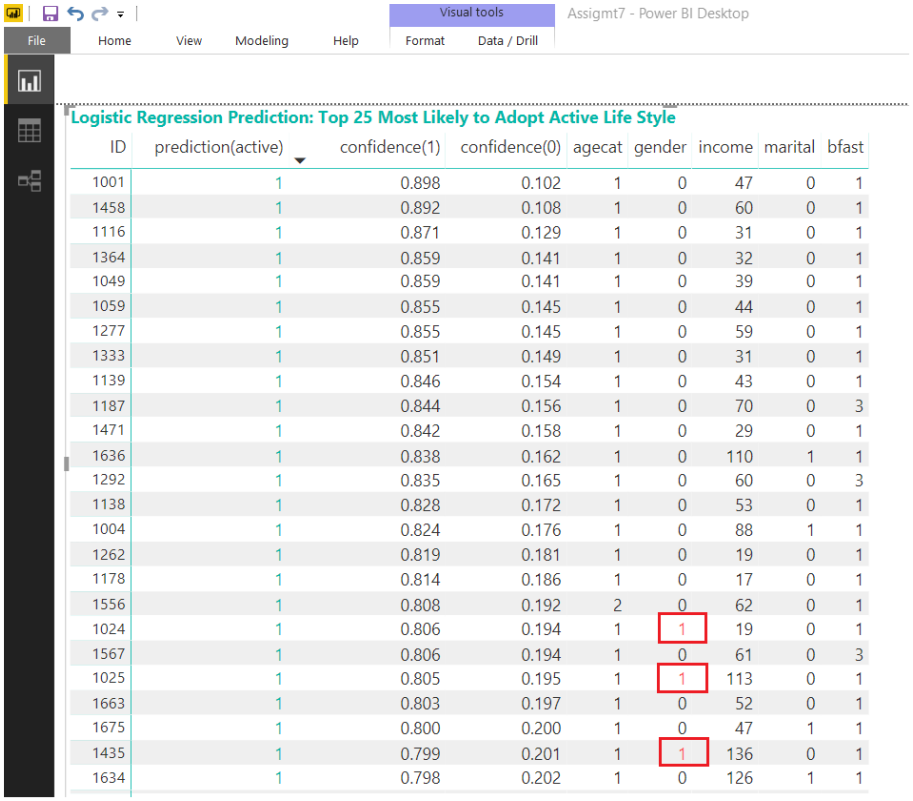
Figure 12. Top 25 most likely to adopt active life has all but one in the under 30 age category



Logistic Regression Prediction: Top 25 Most Likely to Adopt Active Life Style

ID	prediction(active)	confidence(1)	confidence(0)	agecat	gender	income	marital	bfast
1001	1	0.898	0.102	1	0	47	0	1
1458	1	0.892	0.108	1	0	60	0	1
1116	1	0.871	0.129	1	0	31	0	1
1364	1	0.859	0.141	1	0	32	0	1
1049	1	0.859	0.141	1	0	39	0	1
1059	1	0.855	0.145	1	0	44	0	1
1277	1	0.855	0.145	1	0	59	0	1
1333	1	0.851	0.149	1	0	31	0	1
1139	1	0.846	0.154	1	0	43	0	1
1187	1	0.844	0.156	1	0	70	0	3
1471	1	0.842	0.158	1	0	29	0	1
1636	1	0.838	0.162	1	0	110	1	1
1292	1	0.835	0.165	1	0	60	0	3
1138	1	0.828	0.172	1	0	53	0	1
1004	1	0.824	0.176	1	0	88	1	1
1262	1	0.819	0.181	1	0	19	0	1
1178	1	0.814	0.186	1	0	17	0	1
1556	1	0.808	0.192	2	0	62	0	1
1024	1	0.806	0.194	1	1	19	0	1
1567	1	0.806	0.194	1	0	61	0	3
1025	1	0.805	0.195	1	1	113	0	1
1663	1	0.803	0.197	1	0	52	0	1
1675	1	0.800	0.200	1	0	47	1	1
1435	1	0.799	0.201	1	1	136	0	1
1634	1	0.798	0.202	1	0	126	1	1

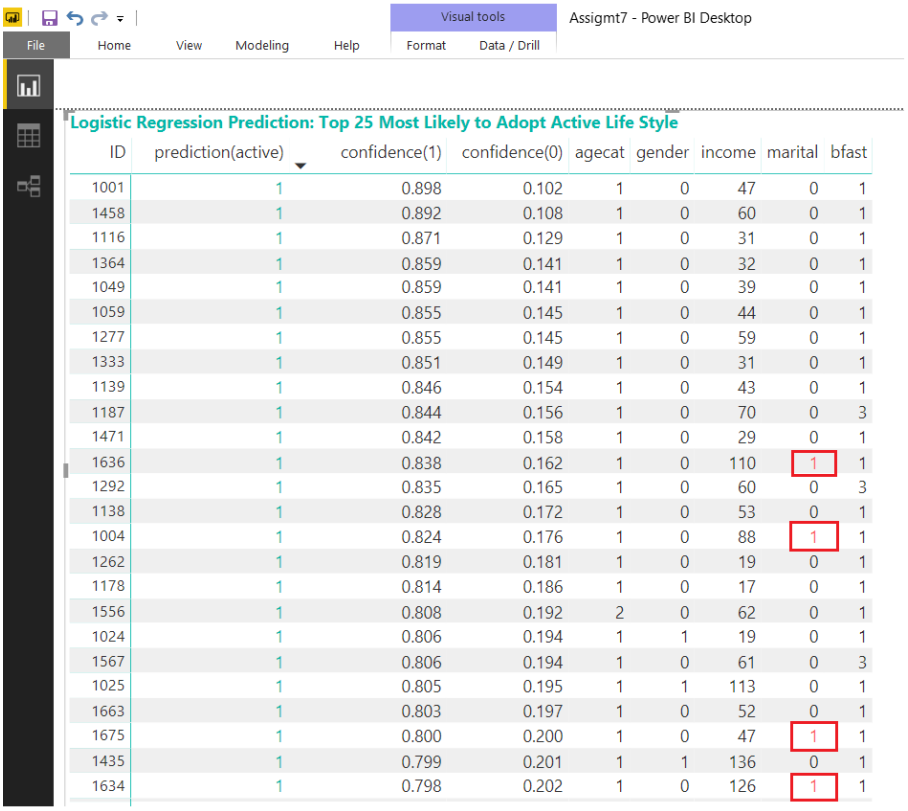
Figure 13. Top 25 most likely to adopt active life has all but 3 females



Logistic Regression Prediction: Top 25 Most Likely to Adopt Active Life Style

ID	prediction(active)	confidence(1)	confidence(0)	agecat	gender	income	marital	bfast
1001	1	0.898	0.102	1	0	47	0	1
1458	1	0.892	0.108	1	0	60	0	1
1116	1	0.871	0.129	1	0	31	0	1
1364	1	0.859	0.141	1	0	32	0	1
1049	1	0.859	0.141	1	0	39	0	1
1059	1	0.855	0.145	1	0	44	0	1
1277	1	0.855	0.145	1	0	59	0	1
1333	1	0.851	0.149	1	0	31	0	1
1139	1	0.846	0.154	1	0	43	0	1
1187	1	0.844	0.156	1	0	70	0	3
1471	1	0.842	0.158	1	0	29	0	1
1636	1	0.838	0.162	1	0	110	1	1
1292	1	0.835	0.165	1	0	60	0	3
1138	1	0.828	0.172	1	0	53	0	1
1004	1	0.824	0.176	1	0	88	1	1
1262	1	0.819	0.181	1	0	19	0	1
1178	1	0.814	0.186	1	0	17	0	1
1556	1	0.808	0.192	2	0	62	0	1
1024	1	0.806	0.194	1	1	19	0	1
1567	1	0.806	0.194	1	0	61	0	3
1025	1	0.805	0.195	1	1	113	0	1
1663	1	0.803	0.197	1	0	52	0	1
1675	1	0.800	0.200	1	0	47	1	1
1435	1	0.799	0.201	1	1	136	0	1
1634	1	0.798	0.202	1	0	126	1	1

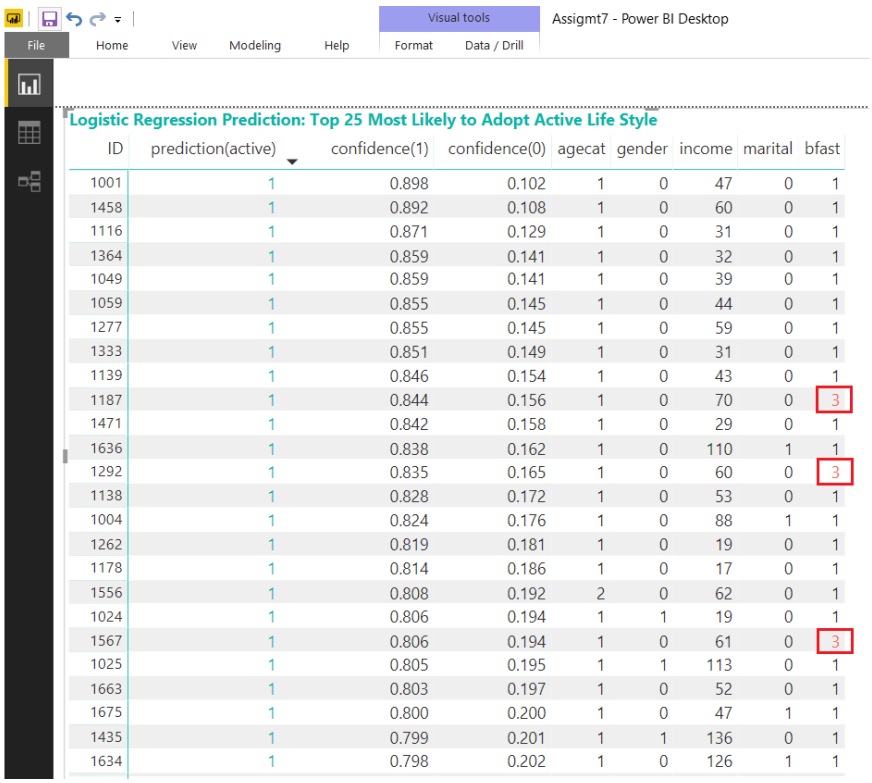
Figure 14. Top 25 most likely to adopt active life has all but 4 singles for marital attribute



Logistic Regression Prediction: Top 25 Most Likely to Adopt Active Life Style

ID	prediction(active)	confidence(1)	confidence(0)	agecat	gender	income	marital	bfast
1001	1	0.898	0.102	1	0	47	0	1
1458	1	0.892	0.108	1	0	60	0	1
1116	1	0.871	0.129	1	0	31	0	1
1364	1	0.859	0.141	1	0	32	0	1
1049	1	0.859	0.141	1	0	39	0	1
1059	1	0.855	0.145	1	0	44	0	1
1277	1	0.855	0.145	1	0	59	0	1
1333	1	0.851	0.149	1	0	31	0	1
1139	1	0.846	0.154	1	0	43	0	1
1187	1	0.844	0.156	1	0	70	0	3
1471	1	0.842	0.158	1	0	29	0	1
1636	1	0.838	0.162	1	0	110	1	1
1292	1	0.835	0.165	1	0	60	0	3
1138	1	0.828	0.172	1	0	53	0	1
1004	1	0.824	0.176	1	0	88	1	1
1262	1	0.819	0.181	1	0	19	0	1
1178	1	0.814	0.186	1	0	17	0	1
1556	1	0.808	0.192	2	0	62	0	1
1024	1	0.806	0.194	1	1	19	0	1
1567	1	0.806	0.194	1	0	61	0	3
1025	1	0.805	0.195	1	1	113	0	1
1663	1	0.803	0.197	1	0	52	0	1
1675	1	0.800	0.200	1	0	47	1	1
1435	1	0.799	0.201	1	1	136	0	1
1634	1	0.798	0.202	1	0	126	1	1

Figure 15. Top 25 most likely to adopt active life has all but 3 that eat breakfast bars



Logistic Regression Prediction: Top 25 Most Likely to Adopt Active Life Style

ID	prediction(active)	confidence(1)	confidence(0)	agecat	gender	income	marital	bfast
1001	1	0.898	0.102	1	0	47	0	1
1458	1	0.892	0.108	1	0	60	0	1
1116	1	0.871	0.129	1	0	31	0	1
1364	1	0.859	0.141	1	0	32	0	1
1049	1	0.859	0.141	1	0	39	0	1
1059	1	0.855	0.145	1	0	44	0	1
1277	1	0.855	0.145	1	0	59	0	1
1333	1	0.851	0.149	1	0	31	0	1
1139	1	0.846	0.154	1	0	43	0	1
1187	1	0.844	0.156	1	0	70	0	3
1471	1	0.842	0.158	1	0	29	0	1
1636	1	0.838	0.162	1	0	110	1	1
1292	1	0.835	0.165	1	0	60	0	3
1138	1	0.828	0.172	1	0	53	0	1
1004	1	0.824	0.176	1	0	88	1	1
1262	1	0.819	0.181	1	0	19	0	1
1178	1	0.814	0.186	1	0	17	0	1
1556	1	0.808	0.192	2	0	62	0	1
1024	1	0.806	0.194	1	1	19	0	1
1567	1	0.806	0.194	1	0	61	0	3
1025	1	0.805	0.195	1	1	113	0	1
1663	1	0.803	0.197	1	0	52	0	1
1675	1	0.800	0.200	1	0	47	1	1
1435	1	0.799	0.201	1	1	136	0	1
1634	1	0.798	0.202	1	0	126	1	1

BUSINES RECOMMENDATIONS

Evaluation of findings suggest that single females under the age of 30 who eat breakfast bars for breakfast are most likely to adopt a healthy life style. So, the report recommends that the NGO reach out to these 25 individuals with their unique ID numbers to start the campaign for an active life style, and recruit these 25 to be leaders/promoters for others who are single, female, relatively younger and likely to adopt an active life style.

The report does recommend that data on additional attributes be collected for the consideration and analysis of active life style, such as smoking, drinking, and exercise. There are data assumptions to begin with and the performance vector showed that the accuracy of prediction could be higher. Additional attributes could yield a more informed prediction, since the existing attributes did not show a very strong inherent correlation.