Project Technical Summary Report
CSC423
Group Name: TBD

1. Abstract

Our goal was to find out if a person thinks he or she is money saver or spender.
The approach was to use logistic regression to find out what have the biggest influence on above perception.
We found out that in general a perception if a person is money spender or saver depends on spending habits like entertainment, appearance, eating out, and in same models on a place of living.
Main limitation of our data set was that for some people saving $100 per month is a lot of money and for some is not. Our data set did not include any information about available budget or any information what a person considered as saving or spending money. (MC)

2. Introduction

The goal of this project is to find out the measurements to determine the survey participants if he or she is a money saver or spender. There can be possible relationship from the wiliness of one person spending on something such as entertainment, appearance to this person's financial behavior. The data we have for analysis is from 'Young People Survey' that contains various topics and the age of participants are from 15 to 30. In this project we only extract the data with potential relationship to people's financial behavior (Young People Survey).
The outcome of this project can be helpful to what elements can be used to identify one person is a money saver or not, and these elements would become reference for people who are trying to save money. For example, if a strong willingness of spending on entertainment will result one as a money spender, then we can assume that limitation of the spending on entertainment can solve the money saving issue.

3. Methodology

Manish Singh section:

Dataset source – https://www.kaggle.com/miroslavsabo/young-people-survey

1. Filtered the original dataset by choosing 14 variables from total of 151 variables.
2. Flagged and deleted missing values (Data cleaning)
3. Finalized Logistic regression to perform analysis.
4. Checked number of observations, sample size
5. Checked for independent and dependent variables.
6. Created dummy and binary variables to perform the regression.
7. Checked for Interaction terms
8. Checked for highest influence variable
9. Checked for significant predictors
10. Wrote the full model equation.
11. Found the Multi-collinearity; Estimated Correlation Matrix.
12. Performed Diagnostics and residual Analysis which included outliers, influence diagnostic graph, and influential points
13. Decided to use to split models making dataset 1 and dataset 2 with different split model and seed
14. Split the data into Training and Testing data (75 - 25) (Deep Analysis)
15. Used Model selection method: Forward, Stepwise
16. Used Model Selection Criteria: R2, AIC, SC, Goodness of Fit Test p-value, standard error.
17. Found the Multi-collinearity; Estimated Correlation Matrix.
18. Performed Diagnostics and residual Analysis which included outliers, influence diagnostic graph, and influential points
19. Tested global null hypothesis
20. Wrote the final model equation using the Maximum likelihood estimates.
21. Analyzed the conditional effect of each variable, odds ration estimates
22. Performed Prediction with Final Model on Training Data using test data to compute phat, lcl, ucl and predicted probability
23. Generate classification table to compute cut-off value.
24. Using cut-off value computed classification matrix
25. Using the classification matrix found out TP, FP, TN, FN
26. Calculated Sensitivity, Accuracy, Precision, Specificity, F-metric
27. Created dataset 2 by Splitting the data into Training and Testing data (60 - 40) (To compare result from Dataset #1)
28. Performed step 19, 20, 21, 22, 23, 24.
29. Compared result from 24 and 26 to find a better model.
30. Keep the better model and discard the other.

Michal Chowaniak section:

Young People Survey - Explore the preferences, interests, habits, opinions, and fears of young people. Data set was obtained from https://www.kaggle.com/miroslavsabo/young-people-survey

1. Chosen 16 out of 151 available variables,
2. Removed tuples with missing data 947 left out of 1011,
3. Prepared boxplots,
4. Checked for one independent binary variable,
5. Checked for number of observations,
6. Checked Sample size,
7. Created dummy variables, create binary variables for categorical variables,
8. Checked for multicollinearity: Estimated Correlation Matrix,
9. Checked for highest influence variables,
10. Checked for significant predictors,
11. Diagnostics and residual analysis:
a. outliers (Pearson Residual or Deviance Residual abs val >= 3), influence diagnostics graph,
b. influential points: Threshold value |Dfbeta|2/sqrt(n) , influence diagnostics graph,
12. Wrote full model,
13. Split dataset into training and testing (proc surveyselect)(use different seed value for each person in a group)
14. Build model with training data set only,
15. Model selection method: Stepwise, Backwards,
16. Model selection criteria: R^2, AIC, SC, Goodness of Fit Test LR, Goodness of Fit Test p-value, predictors selected, standard error of predictors,
17. Checked for multicollinearity: Estimated Correlation Matrix,
18. Diagnostics and residual analysis:
a. outliers (Pearson Residual or Deviance Residual abs val >= 3), influence diagnostics graph,
b. influential points: Threshold value |Dfbeta|2/sqrt(n),
c. influence diagnostics graph,
19. Tested global Null Hypothesis,
20. Wrote down the final model equation (Analysis of Maximum Likelihood Estimates),
21. Analyzed conditional effect of each variable, odds ratio estimates,
22. Using test dataset computed predictions (probability (phat), lower confidence interval (lcl), upper confidence interval (ucl)), predicted probability if Y,
23. Generated classification table to identify cut off value, e.g. 0.5,
24. Using cut-off value of eg. 0.5 classified predicted probability of Y into 1 (over 0.5) or 0 (less or equal 0.5) ,
25. Compared observed Y with predicted Y,
26. Measured performance based on Observed vs Predicted Y: classification matrix,

27. Found out TP, FP, TN, FN,
28. Calculated sensitivity, specificity, accuracy, precision, F-metric,
29. Repeated number 13 to 28 to create a second model,
30. Based on above calculated numbers decided which model was better.

Priyank Beno Cerejo section:

Methodology: -

Obtained dataset from https://www.kaggle.com/miroslavsabo/young-people-survey

1. Exploratory Analysis: Using the raw data selected variables needed for analysis to find if a person saves money or not, i.e. took 14 needed variables from total 151 variables.
2. Data Cleaning: Flagged the missing values, cleaned the data by removing the empty values, i.e. deleted the missing data.
3. Regression Model: Chose Logistic regression to perform analysis.
4. Response Variable: Checked for independent and dependent variables.
5. Dummy Variables: Created dummy and binary variables to perform the regression.
6. Interaction Term: Checked for Interaction Term
7. Full Model Diagnostics: Verified the full model for the whole dataset.. Checked the highest influencing variables.
8. The Model Equation: Wrote the full model.
9. Checking for Outliers, Influential Points and Collinearity
10. Data Splitting: Split the data into Training and Testing data.
11. Checked if the data was split correctly.
12. Verified the full model for the training dataset.
13. Model Selection: Used Model selection method: Stepwise, Forward.
14. Finding Odds Estimates:
15. Using Prediction Model
16. Finding Cut-off Value
17. Computing Classification Matrix
18. Creating 2nd Model: Created 2nd model to find the best fit.
19. Checking Cut-off value
20. Computing Classification Matrix
21. Comparing the Both Models

Rushabh Shah section:

1. Data exploration: explore the data to see if any relationships or trends are present in the data and between variables.
2. Generate scatter plots of each variable against the dependent variable, finances
3. Generate boxplots of each independent variable against dependent variable, finances
4. Data cleaning: remove entire rows of data for rows that have missing values
5. Generate regression model using all variables and data
6. Check for one independent binary variable, and transform/bin variable
7. Create dummy variables for any non-integer variables
8. Generate a full logistic regression model
9. Check to see which variables have the most significance
10. Check outliers, influential points and collinearity
11. Diagnostics and residual analysis
12. Split dataset set into 60/40 training and testing sets
13. Build new logistic regression model with training set
14. Model selection using forward selection
15. Check for collinearity, influence diagnostics graph and outliers
16. Tested global null hypothesis
17. Wrote down final model equation
18. Analyzed odds ratio and estimates of each variable
19. Generate classification table to identify cut-off value
20. Data prediction with training set
21. Generated classification matrix of observed vs. predicted Y values
22. Analyzed predictive performance using predicted probability and 95% confidence limits

Valentine Silvester Correia section:

1. Modified the raw data variables by choosing the needed variables, i.e. took 15 needed variables from total 150 variables.
2. Flagged the missing values, cleaned the data by removing the empty values, i.e. deleted the missing data.
3. Chose Logistic regression to perform analysis.
4. Checked for independent and dependent variables.
5. Created dummy and binary variables to perform the regression.
6. Prepared boxplots for Finances before and after converting it to binary.
7. Prepared Histogram for variable Finances before converting it to binary.
8. Checked for interactions between relevant terms.
9. Verified the full model for the whole dataset.
10. Found the Multi-collinearity; Estimated Correlation Matrix, Model Diagnostics.

11. Checked for Multicollinearity, Influential Point and Outliers.
12. Split the data into Training and Testing data. (First Model)
13. Checked if the data was split correctly.
14. Verified the full model for the training dataset.
15. Used Model selection method: Stepwise.
16. Used Model selection method: Forward.
17. Used Model Selection Criteria: R^2, AIC, SC, Goodness of Fit Test p-value, standard error.
18. Checked prediction data on full model.
19. Wrote full model, computed predicted value on training dataset and obtained the cut-off value.
20. Used cut-off value to compute the classification Matrix.
21. Computed classification matrix.
22. Now for Second Model followed steps 11-21.
23. Compared Model First and Second.

Yesheng Qin section:

1. Data exploration: check and edit the original file and modify it into data with potential relation to identify the finance behavior of one person, check the sample size to see if it has enough samples
2. Data cleaning: clean the dataset such as remove blank value rows in order to make sure the accuracy
3. Choose possible Regression for the goal
4. Check for possible binary variable and choose one that is most likely to be the response variable to determine the finance behavior of one person, if transform is needed then proceed the transform method for the variable
5. Create dummy variables for anyone that has character or string value
6. check if interaction term can be applied
7. Check for the highest significant variables
8. Check outliers, influential points and collinearity
9. Data prediction with given value
10. Splitting data into training and testing set
11. Model selection between stepwise and backward and perform residual analysis on selected model
12. Predictions with training set
13. Comparison and validation
14. Predictive performance
15. Conclusion on the finding and result

4. Analysis, Results and Findings

Manish Singh section:

The dataset we are using is a collection of a survey result conducted over 1010 participants between the age group of 15-30. The survey had a list of 151 questions out of which I have decided to use the below mentioned ones.

<u>Dataset Variables and their description:</u>
Finances: If survey participant is a money saver or spender
Value - Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Shopping_centres: If survey participant like to spend on shopping centers.
Value - Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Branded_clothing: If survey participant prefers branded clothing to non-branded.
Value - Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Entertainment_spending: If survey participant spends a lot of money on partying and socializing.
Value - Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Spending_on_looks: If survey participant spends a lot of money on his/her appearance.
Value - Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Spending_on_gadgets: If survey participant spends a lot of money on gadgets.
Value - Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Spending_on_healthy_eating: If survey participant happily pays more money for good, quality or healthy food.
Value - Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Age: Age of the survey participant (integer)
Siblings: Number of siblings of the participant (integer)
Gender: Gender of participant (female or male)
Education: Highest education achieved by the participant (Currently a Primary school pupil - Primary school - Secondary school – College/bachelor’s degree – master’s degree)
Only_child: If participant is the only child
Value – yes or no
Village_town: If participant spends most time of childhood in village or city
Value - village or city
House_flats: if participant spends most time of childhood in house or block of flats
Value - house or block of flats

As seen above I have a total of 14 variables with 1010 observation.

After filtering out on the final variables to be used from the original dataset I checked if there was a need of any data cleaning i.e. to remove any missing values and I used the manual filters and did find out many missing values.

Data Cleaning

To make the dataset usable for SAS, I used a flag variable and set it to '1' for rows with missing values and later excluded rows having flag value '1'*.*
This step was critical because blank values can have major negative impact on our final result.

The resulting dataset had a total of 986 observations

Regression Model

The agenda of this whole analysis is to answer a 'yes or no' question.
The question I asked was to see if a particular person likes to save money or doesn't like saving money.

As per the principle of data analysis,
Logistic regression can be used for classification purposes, by identifying cases with probability of success *larger than a certain threshold.*

Thus, the model I decided to use was *Logistic regression.*

Response variable and Dummy variables:

Since I chose the logistic regression model, the next step was to decide on a response variable.
I chose the Finances variable as my response variable as that serves the purpose of indicating if the person is a saver or not.

The Finances variable indicates if the survey participant is a money saver or spender.
The values in it are integers ranging from 1 to 5 where 1 is Strongly disagree and 5 is Strongly agree

I have decided to use Finances value (1, 2. 3) as Money spenders and values (4, 5) as money savers.
Hence, I have a new variable d_Finances which will be set to '0' if the Finances value is in (1, 2. 3)
And set to '1' if the Finances value is in (4, 5).

Other Dummy Variables:

Since I noticed a few variables which had non-integer values, I decided to use dummy variable for them. They include gender, only_child, village_town, house_flats, and Education.

Below are the variables along with their original value and the corresponding dummy variable and values.

Original Variable	**Dummy Variable**	**Original Variable Value**	**Corresponding Dummy Variable Value**
Gender	**d_Gender**	**male**	**1**
		female	**0**
Only_child	**d_Only_child**	**Yes**	**1**
		No	**0**
Village_town	**Village_town**	**village**	**1**
		city	**0**
House_flats	**House_flats**	**house/bungalow**	**1**
		block of flats	**0**
Education	**d_Education**	**primary school**	**1**
		secondary school	**2**
		college/bachelor degree	**3**
		masters degree	**4**
		currently a primary school pupil	**5**
		doctorate degree	**6**

<u>Interaction Terms</u>

I thought of the following two scenarios where interaction terms could be created. Refer figure 1

Variables	Why?	P – Value	Action taken
Age and Spending on healthy eating	I think as one grows older from teenage into late 20's, one becomes more cautious as to what he/she eats, and this could be a very interesting variable to consider	0.3329	Discard
Age and Education	Education, in most cases, is directly proportional to Age. Considering this variable sounds interesting as well.	0.3466	Discard

Since, both the interaction variable had a low insignificant p-value and hence discarding them and considering the predictors we already have is the approach I took.

Fitting full model with standard co-efficients

Next, I decided to check the variable for their influence and whether they are significant to predict the odds of finding if the person is a money saver or not.
I did this on the full model before splitting my dataset in test and training sets to have a sense of what might be important beforehand.

Highest influence variable

As seen in the screenshot (Figure 2) of Maximum Likelihood Estimates, Entertainment_spending is the most influential variable followed by Spending_on_looks followed by d_Village_town followed by Spending_on_healthy_eating followed by d_Gender followed by Siblings followed by d_Only_child followed by Age followed by Shopping_centres followed by d_House_flats followed by Branded_clothing followed by Spending_on_gadgets followed by d_Education.
Thus, I can clearly infer here that the spend on entertainment and spend on looks are the two most influential variables.
Variables having Significant Effect
As seen in the screenshot (Figure 2) of Maximum Likelihood Estimates, variables Entertainment_spending, Spending_on_looks, d_Village_town are the only variables which have their 'Pr > ChiSq' greater than 0.05 and they are the significant variables.
The results of influence variable and significant variables are making sense since, all of the most influence variable also have a significant effect on the model.
Full Model Equation
The full logistic regression model to predict probability of d_Finances p=Pr(d_Finances=1) is fitted using
PROC LOGISTIC:

Log(p/(1-p))= 0.1718 – 0.0088 Shopping_centres + 0.00246 Branded_clothing - -0.2777 Entertainment_spending - 0.1817 Spending_on_looks + 0.00231 Spending_on_gadgets + 0.1031 Spending_on_healthy_eating + 0.0112 Age - 0.0401 Siblings - 0.1024 d_Gender + 0.0811 d_Only_child + 0.3996 d_Village_town - 0.0154 d_House_flats - 0.0041 d_Education

where   d_Gender = 1 when Gender = ‘male’
d_Gender = 0 when Gender = ‘female’
d_Only_child = 1 when Only_child="yes"
d_Only_child = 0 when Only_child="no"
d_Village_town = 1 when Village_town="village"
d_Village_town = 0 when Village_town=”city"
d_House_flats = 1 when House_flats="house/bungalow"
d_House_flats = 1 when House_flats=" block of flats"
d_Education = 1 when Education = "primary school"
d_Education = 2 when Education = "secondary school"
d_Education = 3 when Education = "college/bachelor degree"
d_Education = 4 when Education = "masters degree"
d_Education = 5 when Education = "currently a primary school pupil"
d_Education = 6 when Education = "doctorate degree"

<u>Checking full model with multicollinearity influential points and outliers</u>

Multicollinearity

As seen in the correlation table (screenshot was very big to fit) we can clearly see that no two variables have the absolute value over 0.9.

Thus, we can clearly infer than no Multicollinearity issue exists here.

Outliers

As seen in the influence diagnostic screenshot (Figure 3) there are no values with greater than .

Thus, we can clearly say there no outliers present here

Influential Points

Total observation = 986

|Dfbeta| > 2/sqrt (986)

> 0.0636929755298482 approximately = 0.06

When I check the the Dfbeta value against the results obtained, there are many plots having influential points.

But, most of the predictors in our dataset are binary values or values having integer ranging from 1 to 5. I am skipping checking their plots and will concentrate on variables age and siblings as they have more varied values.

For age, our dataset is based on survey based on people of ages 15 to 30. After checking the plot and checking the values in the dataset where the influential points were seen, I don't see any ages that are abrupt. I believe, the influential point in the plot is a result of age predictor along with the other predictors.

Hence, I will ignore the age influential points

For Siblings, on analyzing the variable, I can see that the Siblings value above Dfbeta are either 5, 6 and 10.

I will consider deleting these values.

976/13 = approximately 75

Y = 1 à 338/13 = approximately 26

Y = 0 à 638/13 = approximately 49

Based on the above result we can say that we have enough observations to perform data split.

Data Splitting

To come up with a better model, I decided to use two training datasets using different seed values and split models.

On dataset#1 - I have used the 75-25 model and the seed value of 495857

On dataset#2 – I have used the 60 - 40 model and the seed value of 15865

On dataset #1, I will perform deep analysis of every step mentioned in the methodology. But, I limited the scope of dataset #2 to only check for the end classification matrix and compare it with the matric found from dataset #1. I have made an assumption that within dataset #2, all the analysis ultimately will lead us to the desired result with probably a slight variance but the variances should not impact the validation results of both datasets.

Data Splitting using 75 – 25 model (Detailed Analysis)

After all the above test and confirming that we have enough observations to perform data split into training and testing, I split the data using the seed value 495857 and the 75 – 25 model.

It resulted in a training dataset of 732 observations and the testing dataset of 244 observations.

Model Selection

I have selected the Forward and Stepwise model as my two models for analysis.

Below are the results obtained after analyzing results from both the models, (Figure 4 and 5)

	Models	
	Forward	**Stepwise**
R^2	**0.0518**	**0.0518**

AIC	**905.219**	**905.219**
SC	**923.558**	**923.558**
GOF – LR	**38.5389**	**38.5389**
P- Value	**0.0001**	**0.0001**
Predictors	**Entertainment_spending**	**Entertainment_spending**
	Spending_on_looks	**Spending_on_looks**
	d_Village_town	**d_Village_town**
Predictors and Standard Error	**Entertainment_spending - 0.0737**	**Entertainment_spending - 0.0737**
	Spending_on_looks - 0.0728	**d_Village_town - 0.0728**
	d_Village_town - 0.1728	**Spending_on_looks - 0.1728**

Thus, we can clearly see that the results are exactly same for both models and we can select either of the two and I will go with the stepwise model as my final model.

Also, we can see we are left with only three predictors in the final model and it makes complete sense because these three had the highest influential variable and had the most significant effect.

Checking full model with multicollinearity influential points and outliers

Multicollinearity
As seen in the correlation table screenshot (Figure 6) we can clearly see that no two variables have the absolute value over 0.9.
Thus, we can clearly infer than no Multicollinearity issue exists here.

Outliers
As seen in the influence diagnostic screenshot (Figure 7) there are no values with greater than .
Thus, we can clearly say there no outliers present here
Influential Points
Total observation = 732
|Dfbeta| > 2/sqrt (732)
> 0. 0739221270954573 approximately = 0. 07

When I check the Dfbeta value against the results obtained, there are many plots having influential points.
But, all three predictors (Entertainment_spending, Spending_on_looks, d_Village_town) in our result are binary values or values having integer ranging from 1 to 5. I am skipping checking their plots.

Null Hypothesis

H0: null model b1=b2=0
Ha: Model M1 with b1 and b2 ≠ 0
LR statistic = 38.5389 with chi-square distribution with 3 DF.

Since the P-value is almost zero (0.0001), it clearly indicates that the null hypothesis can be rejected, and current model is better than the null model where at least Entertainment_spending, Spending_on_looks, d_Village_town can explain the probability of a person being a money saver or not.

Final Model Equation

Log(p/(1-p)) = 0.6011 –0.2686 Entertainment_spending -0.1674 Spending_on_looks + 0.4495 d_Village_town

Where d_Village_town = 1 when Village_town="village"
d_Village_town = 0 when Village_town=”city"

Conditional effect of each variable

As seen in the table below, Entertainment_spending and Spending_on_looks have a negative impact on the final result i.e. if one spends on entertainment, the chances of him/her being a saver reduces by 23.6% and with a 95% confidence that the average decrease is between -33.80% and -11.70%
Likewise, if one spends on looks, chances of him/her being a saver reduces by 15.4% and with a 95% confidence that the average decrease is between -26.70% and -2.40%
But, if one stays in a village, the chances of him/her being a saver increases by 56.7% and with a 95% confidence that the average increase is between 11.70% and 120%

The result here makes total sense on a logical level as well, since usually, people spending on entertainment and looks will end up losing money and, in most cases, people living in town will save money. (Figure 8)

	Point Estimate	**95% Wald Confidence Limits**	
Entertainment_spending	**-23.6%**	**-33.80%**	**-11.70%**
Spending_on_looks	**-15.40%**	**-26.70%**	**-2.40%**
d_Village_town	**56.70%**	**11.70%**	**120%**

Prediction with Final Model on Training Data

I used two sets of input values to check the predicted data i.e. probability, LCL and UCL. Below are the results I obtained,

<table>
<tr><th colspan="3">Entertainment_spending = 1, Spending_on_looks = 1, d_Village_town = ‘village’</th></tr>
<tr><td><b>Phat</b></td><td><b>LCL</b></td><td><b>UCL</b></td></tr>
<tr><td rowspan="2"><b><i>0.63208</i></b></td><td><b>0.54453</b></td><td><b>0.71170</b></td></tr>
<tr><td><b><i>72.38%</i></b></td><td><b><i>103.75%</i></b></td></tr>
<tr><th colspan="3">Entertainment_spending = 5, Spending_on_looks = 5, d_Village_town = ‘city’</th></tr>
<tr><td><b>Phat</b></td><td><b>LCL</b></td><td><b>UCL</b></td></tr>
<tr><td rowspan="2"><b><i>0.17177</i></b></td><td><b>0.13160</b></td><td><b>0.22108</b></td></tr>
<tr><td><b><i>14.07%</i></b></td><td><b><i>24.72%</i></b></td></tr>
</table>

Thus, we can clearly infer the possible outcomes from this as mentioned below.

A person strongly disagreeing to spend on entertainment and spend on looks and spends most time in a village has a probability of *0.63208* of being a money saver with a confidence interval of *72.38% and 103.75%*

Whereas, a person strongly agreeing to spend on entertainment and spend on looks and spends most time in a city has a probability of *0.17177* of being a money saver with a confidence interval of *14.07% and 24.72%*

The above results make sense and sounds right logically as well because, if a person refuses to spend on entertainment and spend on looks and spends most time in a village, most times he should end up saving money and contrarily,
if a person wants to spend on entertainment and spend on looks and spends most time in a city, most times he should end up not saving money.
(Figure 9)

Calculating cut-off and computing classification matrix

Using the below table, I calculated the cut-off value which is 0.35.
To compute the classification matrix, after getting the cut-off value I compared it to the phat value and accordingly if the phat value was greater than cut-off value (0.35), I set the predictor y variable to 1.
If the phat value was not greater than cut-off value (0.35), I set the predictor y variable to null.

Prob Level	**Sensitivity**	**Specificity**	
0.35	**59.6**	**62.1**	**121.7**

(Full table in Figure 10)

Analyzing classification matrix to find TP, FP, TN, FN

As per the classification matric (Figure 11), the values are
TP = 48, FP = 35, TN = 93, FN = 68

Sensitivity = 48 / (48 + 68) = 0.4137

Accuracy = (48 + 93) / (48 + 93 + 68 + 35) = 0.5778

Precision = 48 / (48 + 35) = 0.5783

Specificity = 93 / (93 + 35) = 0.7266

F-metric = 2(0.5783 * 0.4137) / (0.5783 + 0.4137) = 2*0.2412 = 0.4824

<u>Data Splitting using 60 – 40 model (To compare result from Dataset #1)</u>

Calculating cut-off and computing classification matrix

Using the below table, I calculated the cut-off value which is 0.3.
To compute the classification matrix, after getting the cut-off value I compared it to the phat value and accordingly if the phat value was greater than cut-off value (0.3), I set the predictor y variable to 1.
If the phat value was not greater than cut-off value (0.3), I set the predictor y variable to null.

Prob Level	**Sensitivity**	**Specificity**	
0.3	**79.6**	**41.6**	**121.2**

(Full table in Figure 13)

Analyzing classification matrix to find TP, FP, TN, FN

As per the classification matric (Figure 12), the values are
TP = 95, FP = 32, TN = 99, FN = 164

Sensitivity* = 95/ (95+ 164*) = 0.3667

Accuracy* = (95+ 99*) / (99 + 95 + 164 + 32) = 0.4974

Precision* = 95/ (95+ 32*) = 0.7480

Specificity* = 99 / (99+ 32*) = 0.7557

F-metric = 2(0.7480 * 0.3667) / (0.7480 + 0.3667) = 2*0.2461 = 0.4922

Compare Model from Dataset #1 to Model from Dataset #2

	Model 1	**Model 2**	**Which is better?**
Sample rate	**75/25**	**60/40**	**-**
Seed	**495857**	**15865**	**-**
Model used	**Forward/Stepwise**	**Forward/Stepwise**	**-**
Cut-off value	**0.35**	**0.3**	**-**
Sensitivity	***0.4137***	***0.3667***	***Model 1***
Accuracy	***0.5778***	***0.4974***	***Model 1***

Precision	***0.5783***	***0.7480***	***Model 2***
Specificity	***0.7266***	***0.7557***	***Model 2***
F-metric	***0.4824***	***0.4922***	***Model 2***

As seen in the above table, we can clearly infer than since model 2 has better Precision, Specificity, F-metric Model 2 using the Dataset #2 (60/40) is the better model.

Now, I can perform in-depth analysis on Model 2 and compare again with Model 1 and make a final decision on which model is better.

Michal Chowaniak section:

Our group's dataset was created by Miroslav Sabo, based on a survey of 1010 participants age of 15 to 30, done in 2013 in Slovakia. The survey asked 151 questions from several areas like music, movie, hobbies preferences, phobias, health habits, personality and the most important for our group spending habits and demographics. We decided to focus on following question:

I save all the money I can.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)

Our goal was to build a model using logistic regression, which could predict if a person thinks about himself/herself as a money spender or saver. We used 16 variables from spending habits and demographics sections, which are answers to flowing questions:

1. I enjoy going to large shopping centres.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
2. I prefer branded clothing to non branded.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
3. I spend a lot of money on partying and socializing.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
4. I spend a lot of money on my appearance.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
5. I spend a lot of money on gadgets.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
6. I will happily pay more money for good, quality or healthy food.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
7. Age: (integer)

8. Height: (integer)
9. Weight: (integer)
10. How many siblings do you have?: (integer)
11. Gender: Female - Male (categorical)
12. I am: Left handed - Right handed (categorical)
13. Highest education achieved: Currently a Primary school pupil - Primary school - Secondary school - College/Bachelor degree (categorical)
14. I am the only child: No - Yes (categorical)
15. I spent most of my childhood in a: City - village (categorical)
16. I lived most of my childhood in a: house/bungalow - block of flats (categorical)

Data cleanup and processing included removing 64 out of 1011 observations, because of missing data. My DV was finances (I save all the money I can), which 5 categories were binned in to 2, so logistic regression could be applied. I also binned to binary remaining categorical variables [1,2,3] . Sample size for Y_0 was 38, and Y_1 20. Only height and height*weight variable were removed, because of high correlation to finances [4] . Exploratory analysis of independent variables suggested that there were 4 variables which had high influence and were significant at the same time [5,6] . Those variables were following:

B_ent - I spend a lot of money on partying and socializing: 0-no, 1-yes
B_look - I spend a lot of money on my appearance: 0-no, 1-yes
B_town - I spent most of my childhood in a: City - 0, village - 1
B_eat- I will happily pay more money for good, quality or healthy food: 0-no, 1-yes;

My data set did not have outliers [7] meaning no observation with Pearson Residual over absolute value of 3. Threshold value for influential point was 0.065, but because there was no outliers I decided not to remove influential points [8,9,10,11,12] . The full model was the following [13] .

Log(b_fin = 1/ b_fin =0) = -0.3163 + 0.0127*b_shop + 0.1247*b_clo – 0.5820*b_ent – 0.5332*b_look + 0.0937*b_gadg + 0.3124*b_eat + 0.0454*b_gender – 0.1129*b_hand + 0.00638*b_edu + 0.1186*b_child + 0.4420*b_town – 0.0106*b_house + 0.0332*age – 0.0126* weight – 0.0541*siblings

From this point I run 2 models A and B [29]. Model A had split 80/20 and model B 70/30[14] between training and test data. Both had different seed values. At the end Model B was a little better so I decided to focused on it. I run model selection using stepwise and backward and both came exactly the same [15] . The Final model included the same variables as in exploratory analysis [16, 17]:

Log(p/(1-p) = -0.5925 – 0.5726*b_ent – 0.6075*b_look + 0.4138*b_eat + 0.4318*b_town

I double check for collinearity, outliers [18], influential point and result were similar as previously.

|Dfbeta|> 2/sqrt(n) > 2/sqrt(663) > 2/25.74879 > 0.077674. Threshold was 0.078, decided to keep influential points, no outliers.

After testing global Null Hypothesis [19] my conclusion was to reject H_0 because at least b_ent b_look b-eat b_town were significant and could predict b_fin_train variable.

Odds ratio estimates for IV were as follows [20]:
B_ent (I spend a lot of money on partying and socializing: 0-no, 1-yes;) [21].
If someone starts spending a lot of money on partying and socializing a chance of a person saying I save all the money I can decreases by 43.60 %, with 95% confidence that the average change is between 18.70% and 60.90%

B_look (I spend a lot of money on my appearance: 0-no, 1-yes;) [22]
If someone starts spending a lot of money on appearance a chance of a person saying I save all the money I can decreases by 45.50%, with 95% confidence that the average change is between 20.10% and 62.90%

B_eat (I will happily pay more money for good, quality or healthy food: 0-no, 1-yes) [23]
If someone starts spending money on eating healthy food a chance of a person saying I save all the money I can increases by 51.30%, with 95% confidence that the average change is between 7.40% and 113.00%

B_town (I spent most of my childhood in a: City - 0, village - 1) [24]
If someone moves from a village to a city a chance of a person saying I save all the money I can increases by 54.00%, with 95% confidence that the average change is between 7.30% and 121.00%

I calculated 2 prediction which are in line with commons sense.

Example 1. Predicted probability for a 21 year old female, whose height is 155 cm, whose weight is 80kg, who is right handed, has college degree, grow up in a city, lives in a house, who likes shopping, spending money on clothing, entertainment, appearance, buying healthy food, who does not like buying gadgets is p=0.17 with 95% prediction interval of (0.13, 0.23). The odds of being a money saver for above person are between 13.91 and 25.95[25].

Example 2. Predicted probability for 30 year old male, whose height is 163 cm, whose weight is 100kg, who is left handed, has secondary level education, who grow up in a village, who lives in apartment, who does not like shopping, spending money on clothing, entertainment, appearance, gadgets, buying healthy food is p=0.46 with 95% prediction interval of (0.39, 0.53). The odds of being a money saver for above person are between 47.97 and 71.23[25].

To test performance of my model I first calculated cut off value, which equal to 0.4[26], using classification tables (Figure 26). Than I classified predicted probability in to two groups below and above 0.4 cut off value and created observed vs predicted y variables matrix[27]. My TN (0,0) = 125, TP (1,1) = 42, FN (1,0) = 64, FP (0,1) = 52. Based on that information I calculated sensitivity, accuracy, precision, specificity and f-metric[28]. Only specificity measure was high and equal to 0.70.

High specificity suggested that my model could better predict if a person is NOT saving money[29], and the model is worse in predicting if a person is saving money. This model would be helpful for a company or non-profit organization interested in educating people about benefits of saving money or targeting people with product advertisements, because they spend money instead of saving them.

Priyank Beno Cerejo section

1. Exploratory Analysis

After finding the dataset selected 14 most important variables that will determine the weather a person saves money or not based on a questionnaire. There are values ranging from 1-5 where 1 being the strongly disagree and 5 being strongly agree.
I have selected 14 variables with 1010 observations

Selected Data Variables
Finances: If survey participant is a money saver or spender
Value - Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Shopping_centres: If survey participant like to spend on shopping centers.
Value - Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Branded_clothing: If survey participant prefers branded clothing to non-branded.
Value - Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Entertainment_spending: If survey participant spends a lot of money on partying and socializing.
Value - Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Spending_on_looks: If survey participant spends a lot of money on his/her appearance.
Value - Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Spending_on_gadgets: If survey participant spends a lot of money on gadgets.

Value - Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Spending_on_healthy_eating: If survey participant happily pays more money for good, quality or healthy food.
Value - Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Age: Age of the survey participant (integer)
Siblings: Number of siblings of the participant (integer)
Gender: Gender of participant (female or male)
Education: Highest education achieved by the participant (Currently a Primary school pupil - Primary school - Secondary school – College/bachelor's degree – master's degree)
Only_child: If participant is the only child
Value – yes or no
Village_town: If participant spends most time of childhood in village or city
Value - village or city
House_flats: if participant spends most time of childhood in house or block of flats
Value - house or block of flats (Fig 1.)

2. Data Cleaning

There were many missing values in the dataset and to avoid having a biased analysis I had to remove the missing values. In order to remove the missing values, I first flagged the missing values using an array function and the rows that were not flagged were saved in Need Dataset and the values that were flagged was saved in Original Dataset. The Need dataset contains 963 values after deleting the missing values. Fig 2.

3. Regression Model

Since we want to find the probability of an event weather a person is a spender or a saver I'll be going ahead with Logistic regression.(Fig 3.)

4. Response Variable and Dummy Variables

As per the questionnaire Finance Variable was the one that gave a complete idea about the financial status of the person based on the other variables selected in my dataset. But those values were based on a questionnaire. So I selected finance variable as a response variable.
The finance variable had the values from 1-5 where 1 being the person saves Money and 5 being the person who doesn't save money.
It order to carry out logistic regression model we need to have a binary value for the finance variable. So, in order to create a binary variable, I grouped them as (1,2,3) as 0 and (4,5) as 1.
Dummy Variables:
For logistic regression we need to convert the character values to numeric values. The dummy variables will be as follows.

Gender = New_Gender
L_R_handed= New_L_R_handed
Education= New_Education
Only_child= New_Only_child

Village_town= New_Village_town
House_flats= New_House_flats

5. Interaction Term

Checked for interaction term between Age - Village_Town and Age – Spending_on_healthy_eating. The P-Value was 0.1016 and 0.0646 respectively. Hence we can discard. (Fig 4.)

6. Full Model Diagnostics

Running a full Model diagnostic before splitting the data into Training and Testing to check the standard co-efficient value, Influence points.

To find highest influence variables

As per the screen shot (Apendix) we can see the following highest influence variables.

Entertainment_spendig

Spending_on_looks

New_Village_town

Spending_on_healthy_

New_Only_child

New_Education

Age

New_House_flats

Branded_clothing

Shopping_centres

Spending_on_gadgets

New_L_R_handed

New_Gender

Variable having significant effect

As per the screen shot (Fig 3.) we can see the following variables having significant effect where

PR > ChiSq' having values lesser than 0.05.

Entertainment_spendi

Spending_on_looks

New_Village_town

7. The Model Equation will be

Log(p/(1-p)) = 0.1353 -0.00266 Shopping_centres -0.00285 Branded_clothing -0.2827 Entertainment_spendi -0.1832 Spending_on_looks -0.00158 Spending_on_gadgets + 0.1006 Spending_on_healthy_ + 0.00796 Age -0.0558 New_Gender -0.0653 New_L_R_handed + 0.1529 New_Only_child + 0.3683 New_Village_town + 0.0177 New_House_flats + 0.0439 New_Education

where New_Gender = 1 when Gender = ‘male’
New_Gender = 0 when Gender = ‘female’
New_Only_child = 1 when Only_child="yes"
New_Only_child = 0 when Only_child="no"
New_Village_town = 1 when Village_town="village"
New_Village_town = 0 when Village_town=”city"
New_House_flats = 1 when House_flats="house/bungalow"
New_House_flats = 1 when House_flats=" block of flats"
New_Education = 1 when Education = "primary school"
New_Education = 2 when Education = "secondary school"
New_Education = 3 when Education = "college/bachelor degree"
New_Education = 4 when Education = "masters degree"
New_Education = 5 when Education = "currently a primary school pupil"
New_Education = 6 when Education = "doctorate degree"

8. Checking for Outliers, Influential Points and Collinearity

Outliers

As we can see there are no values greater than + 3. So we can conclude that there are no outliers.(Fig 5.)

Collinearity

As we can see that the absolute value is not greater than 90% between two variables. So we can conclude that there is no collinearity issue.

Influential Points

To check for influential points, we use the formula |Dfbeta| > 2/sqrt (963) = 0.06
But since most of the variables are between 1 to 5 we can skip the influential points. (Fig 5.)

9. Data Splitting

Split the dataset in 75/25 to divide it into training and testing set. The seed value used was 495857 and sample rate to 0.75. The training set has 741 values and the testing set has 247 values. (Fig 6.)

10. Model Selection

Stepwise and Backward Model

After running stepwise and backward selection model I found out that the R-square, AIC, SC, GOF LR, P-Value, Predictor, Standard Error values for both the model are the same. SO we can chose whichever model we want.

Residual Analysis

The R-Square value is as low as 0.0659. Which means it is not a good model. Also, we can see from the likelihood estimates that a person spends a lot on Entertainment and looks.

(Fig 7)(Fig 8)

	Stepwise	Backward
R-Square	0.0474	0.0474
AIC	886.181	886.181
SC	904.453	904.453
GOF LR	34.5473	34.5473
P-Value	<.0001	<.0001
Predictor	Entertainment_spending	Entertainment_spending
	Spending_on_looks	Spending_on_looks
	New_Village_town	New_Village_town
Standard Error	Entertainment_spending = 0.0756	Entertainment_spending = 0.0756
	Spending_on_looks = 0.0726	Spending_on_looks = 0.0726
	New_Village_town = 0.1772	New_Village_town = 0.1772

11. Odds Estimates

From the odds estimates we can calculate that if you live in a village the chances of you SAVING on Entertainment and looks increases by 50.2%

(1.502-1)*100

Also, if a person spend on looks the chances of a person saving money reduces by 20.8%

(0.792-1) *100
and similarly, if a person spends on Entertainment the chances of saving money reduces by 26.8%
(0.732-1) *100

12. Final Model Equation

Log(p/(1-p)) = 0.9445 – 0.3117 Entertainment_Spending – 0.2337 Spending_on_Looks + 0.4066 New_Village_town.

13. Using Prediction Model

Imputing data lines in the final model to find Phat, LCL, UCL
Entertainment_Spending = 1
Spending_on_Looks = 2
New_Village_town = village
Phat= 0.63607
LCL= 0.53750 = [Exp(0.53750)-1]* 100 = 62.97%
UCL=0.72440 = [Exp(0.53750)-1]* 100 = 75.91%

Entertainment_Spending = 5
Spending_on_Looks = 4
New_Village_town = city
Phat= 0.17577
LCL= 0.13314 = [Exp(0. 13314)-1]* 100 = 42.02%
UCL=0.22846 = [Exp(0. 22846)-1]* 100 = 46.23%
(Fig 9.)

From this we can infer that a person who spends most of the time in the village and does not spend much on entertainment and Looks has the higher probability of being a money saver with a lower confidence limit of **62.97%** to Upper confidence limit of **75.91%**
Similarly, A person who spends most of the time in the city and spends on entertainment and looks and lower probability of being a money Saver with a lower confidence limit of **42.02%** to upper confidence limit of **46.23%**

14. Finding Cut-off Value

Prob Level	**Sensitivity**	**Specificity**	
0.1	100	0	100
0.15	97.3	6	103.3

0.2	94.5	12.6	107.1
0.25	85.2	30.9	116.1
0.3	68.4	43.3	111.7
0.35	53.9	62.1	116
0.4	45.7	75.9	121.6
0.45	31.6	87	118.6
0.5	19.5	92.6	112.1
0.55	14.5	95.7	110.2
0.6	4.7	98.6	103.3

15. Computing Classification Matrix

(Fig 11.)
TP= 41
TN=104
FP=46
FN=56

From the classification matrix we found out the following

Sensitivity = 0.422
Specificity = 0.693
Accuracy = 0.587
Precision = 0.471
F-Metric = 0.445

16. Creating 2nd Model

Created a 2nd Model with the sample rate of 60%
Split the data into training and test sets - 60/40;
samprate = 60% of observations to be randomly selected for training set
out = train defines new sas dataset for training/test sets;

17. Checking Cut-off value

Classification Table		

Prob	**Correct**		**Incorrect**		**Percentages**						
Level	**Event**	**Non-**	**Event**	**Non-**	**Correct**	**Sensi-**	**Speci-**	**FALSE**	**FALSE**		
		Event		**Event**		**tivity**	**ficity**	**POS**	**NEG**		
0.1	200	0	393	0	33.7	100	0	66.3	.		100
0.15	195	29	364	5	37.8	97.5	7.4	65.1	14.7		104.9
0.2	186	72	321	14	43.5	93	18.3	63.3	16.3		111.3
0.25	161	146	247	39	51.8	80.5	37.2	60.5	21.1		117.7
0.3	139	202	191	61	57.5	69.5	51.4	57.9	23.2		120.9
0.35	112	256	137	88	62.1	56	65.1	55	25.6		121.1
0.4	84	293	100	116	63.6	42	74.6	54.3	28.4		116.6
0.45	61	334	59	139	66.6	30.5	85	49.2	29.4		115.5
0.5	47	358	35	153	68.3	23.5	91.1	42.7	29.9		114.6
0.55	34	375	18	166	69	17	95.4	34.6	30.7		112.4
0.6	12	382	11	188	66.4	6	97.2	47.8	33		103.2

18. Computing Classification Matrix

(Fig 12.)
TP= 74
TN=157

FP=69
FN=95

From the classification matrix we found out the following

Sensitivity = 0.4378
Specificity = 0.694
Accuracy = 0.584
Precision = 0.517
F-Metric = 0.47411

19. Comparing the Both Models

	1st Model	**2nd Model**	
Sample Rate	75/25	60/40	
Seed	495857	587634	
Model Selection	Stepwise – Backward	Stepwise – Backward	
Cut of Value	0.4	0.35	
Sensitivity	0. 422	0. 4378	**2nd Model**
Specificity	0. 693	0. 694	**2nd Model**
Accuracy	0. 587	0. 584	**1st Model**
Precision	0. 471	0. 517	**2nd Model**
F-Metric	0. 445	0. 47411	**2nd Model**
Better Model		**2nd Model**	

Sensitivity = 0.422
Specificity = 0.693
Accuracy = 0.587
Precision = 0.471
F-Metric = 0.445

Rushabh Shah section:

Original dataset with 14 variables and 1010 observations:

- **Finances**: I save as much money as I can.: Strongly disagree 1-2-3-4-5 Strongly agree (integer).
- **Shopping_centres**: I enjoy going to large shopping malls.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
- **Branded_clothing**: I prefer branded clothing over non-branded clothing.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
- **Entertainment_spending**: I spend a lot of money on personal entertainment.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
- **Spending_on_looks:** I spend a lot of money on my appearance.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
- **Spending_on_gadgets**: I spend a lot of money on electronics and gadgets.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
- **Spending_on_healthy_eating**: I will gladly pay more for good, quality or healthy food.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
- **Age:** age of the survey participant (integer)
- **Number_of_siblings**: how many siblings do I have?: (integer)
- **Gender:** Gender of participant (female or male)
- **Education**: what is the highest level of education I have completed.: (Currently a Primary school pupil - Primary school - Secondary school – College/Bachelor degree – Masters degree)
- **Only_child**: if participant is only child (yes or no)
- **VorC**: I spent most of my childhood in.: (village or city)
- **HorB**: I spent most of my childhood in.: (house or block of flats)

Data cleaning:

With some rows having missing values for different variables, I couldn't use the dataset as-is, so I had to remove each row of data that had missing values at any variable. If I hadn't addressed the missing values, the accuracy of our model might be negatively affected. This cleaning process left the dataset with 972 observations. The code that I used to remove the missing values was: if cmiss(of _character_) + nmiss(of _numeric_) > 0 then delete;

In addition to removing the rows of observations with missing values, I binned the Finances variable data. For the Finances variable, values of 1, 2, and 3 were all binned into 0, meaning all of the individuals who answered with 1, 2, or 3 all were not considered money savers. The values 4 and 5 were binned into 1, meaning everyone who answered with a 4 or 5 was considered a money saver. Along with this, I created dummy variables for those variables that had categorical data, including the variables Gender, Education, Only_Child, VorC and HorB.

- Gender -> sex: 1 = male; 0 = female
- Education -> schooling: 0 = currently in/only primary school; 1 = secondary school; 2 = college/bachelor degree; 3 = masters degree
- Only_child -> only: 1 = yes; 0 = no

- VorC -> Urban: 1 = city; 0 = village
- HorB -> Housing: 1 = house; 0 = block of flats

Regression Model:
Since the goal of our analysis is to create a model that can identify whether a person is a money saver or money spender, I conducted a logistic regression, which required me to bin the values for the finance variable into 1 for saver and 0 for spender. By binning the finances variable into 0 or 1, I would be able to conduct a regression analysis that better matched the results we were trying to predict, which was, is a person a money saver, or a spender. Doing a linear regression analysis would have been useful if we wanted to predict whether the person would score a 1, 2, 3, 4 or 5 for finances, but not if wanted only a 1 or a 0.

Likelihood Estimates table:

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-2.7857	1.7758	2.4610	0.1167
Housing	1	0.0312	0.1787	0.0305	0.8614
Urban	1	0.3841	0.1903	4.0737	0.0436
schooling	1	0.0297	0.1347	0.0487	0.8254
Only	1	-0.0655	0.1832	0.1279	0.7206
Handed	1	-0.1350	0.2315	0.3403	0.5597
sex	1	-0.1232	0.2145	0.3300	0.5656
Number_of_siblings	1	0.0598	0.0795	0.5669	0.4515
Weight	1	0.00414	0.00815	0.2580	0.6115
Height	1	0.0140	0.0107	1.7015	0.1921
Age	1	-0.0302	0.0297	1.0319	0.3097

Spending_on_healthy_	**1**	**-0.1274**	**0.0680**	**3.5125**	**0.0609**
Spending_on_gadgets	**1**	**0.00199**	**0.0641**	**0.0010**	**0.9753**
Spending_on_looks	**1**	**0.1910**	**0.0772**	**6.1198**	**0.0134**
Entertainment_spendi	**1**	**0.2698**	**0.0687**	**15.4397**	**<.0001**
Branded_clothing	**1**	**-0.0118**	**0.0648**	**0.0331**	**0.8557**
Shopping_centres	**1**	**0.0130**	**0.0636**	**0.0417**	**0.8383**

According to the likelihoods estimate table, the variables Urban (whether childhood was spend in village or city) and Entertainment_spending are the highest influence variables.

Outliers and Influential Points:
According to the dataset analysis, there are not outliers or influential points. The influence diagnostics graphs show us that none of the data points have a pearson residual value outside of the range of -2 to 2, which indicates that we have no outliers in our data. Looking at the graph of Dfbetas, we see that their are influential points in our data. The threshold for Dfbetas is 2/sqrt(972), which comes out 0.064, and our graphs clearly show us that we have values beyond this threshold. These influential points are hard to identify and exclude from our dataset because most of our variables have values between 1 and 5, so to remove influential points, we would have to remove every response with a value of 5, seeing as these would be the “influencers”.

Predictions with Full Model:
For the full model, I used Housing Urban schooling Only Handed sex Number_of_siblings Weight Height Age Spending_on_healthy_eating Spending_on_gadgets Spending_on_looks Entertainment_spending Branded_clothing Shopping_centres, with the corresponding values: 0 1 3 0 0 0 2 58 163 19 2 5 2 4 1 4 to make predictions. These values gave us a phat values 0.28196, an lcl of 0.22594, and an ucl of 0.34567. For the second prediction, the corresponding values are 0 1 2 0 0 1 1 77 186 20 4 4 1 3 3 3. These values gave a phat of 0.34987, an lcl of 0.29668 and ucl of .40707. The phat values tell us that our model is not that great at predicting whether an individual is a money saver (finances (event =’1’)).

Data Splitting:
The seed value that I used for my training set is 12345 and a sample rate of 0.60. The training set has 581 samples, and is stored under the variable ‘train’ variable, and the test set has 391 samples, and is stored under the ‘validation’ variable.

Model Selection:
Model selection was done using forward selection. This forward selection had an AIC value of 744.82, and an SC value of 749.187 on the training set, and an AIC value of 561.679 and SC value of 565.734 on the test set. Seeing the AIC value decrease for the test set, as compared to on the training set, tells us that our model performance was fairly good. The decrease in AIC value tells us that the information loss decrease on the test model, as compared to on the training model, and the more information loss we can reduce, the better our model. Forward selection left our final model with 2 variables, Urban and Entertainment_spending. The odds ratio estimate for the variable Urban is 0.744 with a range of 0.484 to 1.144, and for the variable Entertainment_spending is 0.730 with a range of 0.615 to 0.866. The p-values for each of the two variables are extremely low, with the highest p-value being 0.0117. The low p-values tell us that we have a good predictive model for predicting financial behavior, and that we can reject the null hypothesis.

Prediction with Final Model:
Since the final model only has 2 variables, entertainment spending and urban, running predictions on this was a little easier than running predictions on the full model. For this prediction, Entertainment_spending had a value of 1 and Urban has a value of 0, for these values, the phat was 0.57598, with an lcl of 0.45297 and ucl of 0.69024. For the second observation, the values for Entertainment_spending is 2 and for Urban is 1. These values gave us a phat of 0.42446, lcl of 0.35375 and ucl of 0.49842. As compared to the phat values for prediction using the full logistic model, the phat values for prediction using the final logistic model are much higher. This increase in phat value tells me that my new ‘final’ model is better at predicting if a person is a money saver (finances (event = ‘1’)).

Conclusion
The results of my regression indicate that knowing whether a person grew up in a village or in the city and knowing their spending habits regarding entertainment can help us predict whether a person generally prefers to save or spend their money. These two variables are reasonable for the age group that was surveyed, those aged 15-30. Generally those in this age range are still young enough to be influenced by their parents, and the things that they grew up with. For example, those who grew up in smaller villages, generally are going to be savers, whereas those who grew up in the city will generally be spenders, mainly because they grew up with access to so many more places or things they can spend their money on. Along with this, knowing a person's’ spending habits regarding entertainment, can be a very good predictor for predicting their financial behavior. Those who spend a lot of money on entertainment, will more than likely not be money savers, but those are conservative about their entertainment spending, will be more conservative in their spending, aka be savers.

Valentine Silvester Correia section:

Variables: -

Finances	if survey participant will save all the money I can: Strongly disagree 1-2-3-4-5 Strongly agree (integer).
Shopping_centres	if survey participant will enjoy going to large shopping center’s: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Branded_clothing	if survey participant will prefer branded clothing to non-branded : Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Entertainment_spending	if survey participant will spend a lot of money on partying and socializing: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Spending _on_looks	if survey participant will spend a lot of money on my appearance: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Spending_on_gadgets	if survey participant will spend a lot of money on gadgets: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Spending on healthy eating	if survey participant will hapilly pay more money for good, quality or healthy food: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Age	age of the survey participant (integer): ranges from 15-30 years

Siblings	How many siblings does participant have? (integer): Values- 0,1,2,3,4,5,6,10
Gender	Gender of participant: Values- female or male
L_R_handed	if participant is left-handed or right-handed: Values- right handed or left handed
Education	highest education achieved by the participant: Values- Primary school, Secondary school, College/Bachelor degree, Masters degree
Only_child	if participant is only child: Values- yes or no
Village_town	if participant spends most time of childhood in village or city: Values- village or city
House_flats	if participant spends most time of childhood in house or block of flats: Values- house or block of flats

Data Cleaning: -

This was a large dataset with missing values and so it was necessary to handle the missing values so as not to affect it negatively. I followed the link https://measuringu.com/handle-missing-data/ to find possible solutions to deal with the missing data and I tried Listwise Deletion. After deleting I saw that approximately 3.5% of the total entries was deleted and so I planned to stick with Listwise Deletion. To perform deletion of missing data, I first flagged the missing values and then deleted the flagged values to obtain the new dataset with 986 values.

Regression Model: -

As we have to predict if a person is a saver or not a saver, we can say this is a dichotomous analysis with Finances being our dependent variable, hence we can use Logistic regression.

Response Variable: -

As proposed in the project proposal, Finances will be my response variable i.e. the dependent variable. I converted the DV i.e. Finance Variable to bin values. Values 1, 2, 3 as “0” i.e. spenders and value 4, 5 as “1” i.e. saver.

Dummy Variables: -

To represent certain variables with distinct categories I have created dummy variable. Below I have listed the dummy variables I have created.

Original Name	Dummy Name	Description
Finances	d_ Finances	
Gender	d_ Gender	male=1; female=0
L_R_handed	d_ L_R_handed	right handed=1; left handed=0
Only_child	d_ Only_child	yes=1; no=0
Village_town	d_ Village_town	village=1; city=0
House_flats	d_ House_flats	house/bungalow=1; block of flats=0
Education	d_ Education	primary school=1; secondary school=2; college/bachelor degree=3; masters degree=4; currently a primary school pupil=5; doctorate degree=6

Fitting Full Model to the data before splitting it to training and testing: -
I moved further to check the variables for highest influential value and find significant predictors to see if they can predict that a person is a saver or not a saver.
I have attached the findings in appendix and from that we can infer that the two most influential variable are Entertainment_spending and Spending__on_looks.
Also, from the table, we can see the variables that have their p-value > 0.05 are Entertainment_spending, Spending_on_looks, d_Village_town and they are the significant predictors.
From observing the highest influential values and significant predictors we can ultimately say that all of the most influential values have a significant effect on the model and that is logical.
I have fitted the full logistic regression model PROC LOGISTIC to predict the probability of d_Finances=Pr(d_Finances=1)
Interaction Term: -
For our analysis we could try interaction between: -

- Age*Education- we can determine the education of a person with respect to the persons age.
- Spending_on_looks*Branded_clothing- we can determine that a person who spends on looks will also tend to spend on branded clothing.

But, after analyzing the p-value for both the terms i.e. 0.4587 and <0.0001 we can say that they are insignificant and therefore we can skip applying the interaction terms for this model.

Likelihood Estimates Table: -

Below is the likelihood estimates table sorted highest influential variable to lowest influential variable: -

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Entertainment_spendi	1	-0.2777	0.0683	16.549	<.0001	0.1823
Spending__on_looks	1	-0.1817	0.0769	5.5859	0.0181	0.121
d_Village_town	1	0.3996	0.1914	4.3609	0.0368	0.1
Spending_on_healthy_	1	0.1031	0.0671	2.3593	0.1245	0.0622
d_Gender	1	-0.1024	0.1592	0.4138	0.52	0.0277
Siblings	1	-0.0401	0.0781	0.2632	0.608	0.0225
d_Only_child	1	0.0811	0.181	0.2005	0.6543	0.0193
Age	1	0.0112	0.0319	0.1227	0.7261	0.0171
Shopping_centres	1	-0.00882	0.0634	0.0194	0.8893	0.00643
d_House_flats	1	-0.0154	0.1793	0.0074	0.9315	0.00417
Branded_clothing	1	0.00246	0.0642	0.0015	0.9694	0.00177
Spending_on_gadgets	1	0.00231	0.0636	0.0013	0.9711	0.00163
d_Education	1	-0.0041	0.1217	0.0011	0.9731	0.00163

Multicollinearity, Influential Points: -

Multicollinearity: -

From the correlation table obtained we see that no 2 variables have the absolute value greater than 0.9 and hence we can say that no Multicollinearity exists.

Influential Points: -

Total observations are 986

|Dfbeta| > 2/sqrt (986)

>0.0636929755298482 approximately = 0.06

After comparing the Dfbeta with the obtained results, we see that there are many plots those have influential points.

Most of the predictors in our dataset are bin i.e. 1 or 0 or integers ranging from 1 to 5.

So, I will here consider taking variable that have varied values and they are Age- ranges from 15 to 30 and Siblings- ranges from 1-6 and 10.

After checking the plot for age I don't see any ages that are unexpected. Hence, we can ignore the age influential points

After checking the plot for Siblings, I can see that the Siblings value above Dfbeta are either 5, 6 and

10 and hence, we will consider deleting these values. So now we have 976 values in total.

Frequency obtained was:

976/13= 75 approx.

Y= 1-> 338/13= 26 approx.

Y= 0-> 638/13= 49 approx.

From the above results we can say that we have enough observations to split the data for traning and testing.

Data Splitting: -

Now I planned to split my data my data for Training and Testing.

Using the random seed value 564786, I divided my training data to 80% i.e. 781 sample size.

And, the remaining 20% observation that is 195 observation I will be using for testing.

Model Selection: -

I planned to go ahead with Stepwise and then Forward model and below are the observation: -

	Models	
	Stepwise	**Forward**
R^2	0.0505	0.0505
AIC	1260.450	1260.450
SC	1265.332	1265.332
GOF – LR	50.4891	50.4891
P- Value	0.0001	0.0001
Predictors	Entertainment_spending	Entertainment_spending
	d_Village_town	d_Village_town
	Spending_on_looks	Spending_on_looks

Predictors and Standard Error	Entertainment_spending à 0.0636	Entertainment_spending à 0.0636
	Spending_on_looks à 0.0627	Spending_on_looks à 0.0627
	d_Village_town à 0.1490	d_Village_town à 0.1490

From the above models we can see that R^2, AIC, SC, etc. all results are exactly same for Stepwise and then forward models. Therefore, we can go ahead with either of the two methods. I plan to go ahead with the Forward model.

Multicollinearity, Influential Points: -

Multicollinearity: -
From the correlation table obtained we see that no 2 variables have the absolute value greater than 0.9 and hence we can say that no Multicollinearity exists.

Influential Points: -
Total observations are 781
|Dfbeta| > 2/sqrt (781)
>0.07156562669 approximately = 0.07
We can skip checking the plots as all our predictors are bin or integers.

Final Model: -

Fitting the final model: -
Log(p/(1-p)) = 0.6054 – 0.2809*Entertainment_spending – 0.1539*Spending_on_looks + 0.3836*d_Village_town.

Prediction Model: -

	Point Estimate	95% Wald Confidence Limits	
Entertainment_spending	-24.8%	-33.6%	-14.8%
Spending_on_looks	-13.7%	-23.7%	-2.4%
d_Village_town	46.9%	9.8%	97.6%

Entertainment_spending = 1, Spending_on_looks = 2, d_Village_town = ‘village’

Phat	LCL	UCL
0.5971	0.5145	0.6745
	67.29%	96.3%
Entertainment_spending = 4, Spending_on_looks = 5, d_Village_town = ‘city’		
Phat	LCL	UCL
0.2161	0.1728	0.2667
	18.87%	30.56%

Cut-off Value: -

Prob level	Sensitivity	Specificity	
0.1	100	0	100
0.15	97.7	0	97.7
0.2	95.8	13.2	109
0.25	86	30.8	116.8
0.3	70.8	47.6	118.4
0.35	53.4	66	119.4
0.4	34.8	77.9	112.7
0.45	25.4	87.6	113
0.5	17.4	92.1	109.5
0.55	5.7	96.5	102.2
0.6	4.9	98.3	103.2

Analyzing classification matrix to find TP, FP, TN, FN: -

TP = 43
FP = 27
TN = 47
FN = 78

<u>Calculating:</u>

Sensitivity = 43/ 43 + 47 = 0.478
Accuracy = 43+78 / 43+78+47+27 = 0.6205
Precision = 43 / 43+27 = 0.614
Specificity = 78 / 78 + 27 = 0.742
F-Metric= 0.537

Second Model: -
Using Backward and Stepwise.
Cut-off Value: -

Prob level	Sensitivity	Specificity	
0.1	100	0	100
0.15	100	0	100
0.2	96.3	5.9	102.2
0.25	86.9	15.5	102.4
0.3	79.5	36.6	116.1
0.35	60.7	59.8	120.5
0.4	41.4	74.8	116.2
0.45	25.8	87	112.8
0.5	15.6	91.1	106.7
0.55	5.7	96.4	102.1
0.6	2.9	98.2	101.1

Analyzing classification matrix to find TP, FP, TN, FN: -

TP = 52
FP = 43
TN = 69
 FN = 128
<u>Calculating:</u>
Sensitivity = 52/ 52+69 = 0.429
Accuracy = 52+128 / 52+128+69+43= 0.616
Precision = 52 / 52+43 = 0.547
Specificity = 128 / 128 + 43 = 0.748
F-Metric= 0.480

Comparing Fist and Second Model: -
Comparing both the models,

	First Model	Second Model	Model Selected
Sample rate	80/20	70/30	
Seed	564786		
Model Selection	Stepwise, Backward		
Sensitivity	0.478	0.429	First Model
Specificity	0.742	0.748	Second Model
Accuracy	0.6205	0.616	First Model
Precision	0.614	0.547	First Model
F-Metric	0.537	0.480	First Model

From the above table, based on the sensitivity, specificity, accuracy, precision and F-metric values we can select the First Model.

From the analysis we can conclude that the person is either money saver or spender based on how much he or she spends on the entertainment and appearance (Entertainment_spending, Spending_on_looks) and where the person spent his/her childhood (d_Village_town).
If a person spent more money on entertainment and look they were spender and they people who grew up in Villages were probably the savers.

Yesheng Qin section:

Original Variables of Dataset

- **Finances**: if survey participant will save all the money I can.: Strongly disagree 1-2-3-4-5 Strongly agree (integer).
- **Shp_centre**: if survey participant will enjoy going to large shopping centres.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
- **Brd_cloth**: if survey participant will prefer branded clothing to non branded.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
- **Enter_spd**: if survey participant will spend a lot of money on partying and socializing.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
- **Spd_lk**: if survey participant will spend a lot of money on my appearance.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
- **Spd_gadgt**: if survey participant will spend a lot of money on gadgets.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
- **Spd_eat**: if survey participant will hapilly pay more money for good, quality or healthy food.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
- **Age:** age of the survey participant (integer)
- **Siblings**: How many siblings does participant have?: (integer)
- **Gender**: Gender of participant (female or male)
- **Education**: highest education achieved by the participant (Currently a Primary school pupil - Primary school - Secondary school – College/Bachelor degree – Masters degree)
- **Only_child**: if participant is only child (yes or no)
- **VorC**: if participant spends most time of childhood in village or city (village or city)
- **HorB**: if participant spends most time of childhood in house or block of flats (house or block of flats)

Total of 14 variables with 1010 observation.

Data cleaning

There are some rows with blank value for some variables, so in order to make my data more accurate I delete the observation with one or more blank values, which gives me a new dataset with 972 observations.
The code for the cleaning section is:

```
if cmiss(of _character_) + nmiss(of _numeric_) > 0 then delete;
```

and it can fetch all rows with either character value or numeric value but blank instead, and delete it for my new dataset.

Regression Model

Because we are trying to identify whether one person is a money saver or a money spender, the best method is to use Logistic Regression to find out the possibility of one event (either a saver or spender). Therefore, we will need to target a response variable that is binary as 1 for saver and 0 for spender.

Response Variable

- **Finances**: if survey participant will save all the money I can.: Strongly disagree 1-2-3-4-5 Strongly agree (integer).

Since this variable is in integer form, it needs to be transformed into binary for logistic model, so I use 'Saver' as the binary variable to indicate either a person is a money saver (Saver=1) or money spender (Saver=0).
Because 'Finances' has 5 kinds of value so I use histogram to check the distribution of each one (see Appendix 'Response variable 'Finances' distribution'):
Value with highest frequency is '3' which may indicate as either spender or saver. I merge value 1,2,3 into money spender and 4,5 into money saver.
Therefore, I have a new variable which is binary to indicate whether the person is money saver or spender:

- **Saver**: if Saver=1 then the person is money saver, if Saver=0 then the person is a money spender

Dummy Variables:

Gender to Gender1 (1,0) , Education to Edu_level(1,2,3,4), Only_child to One_child(1,0), VorC to Village(1,0) and HorB to House(1,0)
See Appendix 'Dummy variable table' for detail

Interaction Term

The most likely interaction terms for this dataset can be Age*Edu_level as the older one person is, he or she may have higher education level, and another one is Spd_lk*Brd_cloth as more serious you are for the appearance, more you may spend on the branded clothes. However, the p-value for both terms are not significant, the first one is 0.2015, and the second one is 0.3953 so we can skip applying the interaction terms for the model.

Likelihood Estimates table

The highest influence variables are 'Enter_spd' which is the spending on entertainment, 'Spd_lk' which is spending on appearance and look and variable 'Village' which is either born in village or city (see Appendix 'Full likelihood estimates table').

Outliers, Influential Points and Collinearity

- Collinearity
- There is no value between two variables having the absolute value greater than 90%, so it can be indicated as no collinearity issue (see Appendix ‘Full correlation table).
- Outliers

Based on the graph (Appendix ‘Full outliers’) that presents the distribution for the residual, there is no plot exceeding absolute value of 3, so it can be indicated as no outlier

- Influential Points

Threshold value = 2/sqrt(n) = 2/31 = 0.065

There are many plots with Dfbeta value higher than the threshold, but most of the variables in this dataset have either binary value or integer value only from 1 to 5, so I would check variables with more flexible values such as Age and Siblings. However, the dataset is from a survey focusing on young people with age from 15-30, so we may consider all the age values are within acceptable range, which leave the Siblings to be checked

According to the graph (Appendix ‘Full influential points’) of influential points we can notice that there is 1 plot above 0.4 and 1 above 0.2. The Siblings value with Dfbeta above 0.4 is 10, which is the largest in the dataset, and the second largest is 6, so we may consider removing it, same as the second one higher than 0.2 which has 6 for the Siblings value, and it increment to 0.28 Dfbeta value after removing the previous one, so we may delete this one as well. the Others are still within the reasonable range and should not affect too much of the data.

The new dataset now has 970 observations in total, so to check the frequency of ‘Saver’ is either 1 or 0, we can do the frequency table for variable ‘Saver’, and use the data to see if we have enough sample to split in to future (see Appendix ‘Full frequency table on Saver’).

970/13 = 75

Y=1 -> 338/13 = 26

Y=0 -> 632/13 = 48.62

Based on the result, we have enough observation for our data splitting step

Predictions with Full Model

I have make two predictions with the full logistic model, the inputs are:

Shp_centre Brd_cloth Enter_spd Spd_lk Spd_gadgt Spd_eat Age Siblings Gender1 Edu_level One_child Village House;

And the first prediction has the given value: 1 2 2 2 1 2 27 2 1 4 0 0 1, the second prediction has: 4 5 3 5 2 4 22 0 0 3 1 1 0, these values are corresponding to the input variables from left to right. The result for the first one is phat=0.38796, lcl=0.25929, ucl=0.53442 and the second one has phat=0.38672, lcl=0.25241 and ucl=0.54079.

The result of these two are not very persuasible to identify which one is more likely to be a money saver as even though the given values are totally different, the result are very similar. Therefore, we need to find out the final model with most significant variables to perform the prediction step.

Data Splitting

The seed value I have for the training set is 438821 and the sample rate at 0.65 (see Appendix ‘Training set table’)

- **Training set**

It has 631 samples in the training set, and I will use another variable ‘train_y’ to represent the selected value of Saver for training set.

- Testing set

Testing set will have the rest samples after deducting the above 679 from training set

Model Selection

- **Stepwise and Backward**

Both stepwise and backward models are having the same value for AIC (816.76), SC (821.21), variables (Enter_spd, Spd_lk), R-square (0.05), etc, so there is no different for choosing either one (see Appendix ‘Stepwise’ and ‘Backward’).

I will use the result from backward for the final model and the variables for the final model are :Enter_spd, spd_lk

- **Residual Analysis**

The odd rations from the estimates table are 0.794 for Enter_spd with range from 0.68 to 0.928, and 0.79 for Spd_lk from 0.677 to 0.921. This indicates that if a person is having their spending level on entertainment or appearance as 1, there will be 0.21 decreased to treat this person as a money saver (see Appendix ‘Odd ratio’).

The correlation table (Appendix ‘Backward correlation table’) gives us the result that no correlation value exceeds 0.9, so there is no collinearity issue for this model. Moreover, the there is no plot over the absolute value of 3 from the diagnostic, so no outliers in the data nor very significant influential points according to the Dfbeta threshold of 0.08 based on the 631 samples in the training set (Appendix ‘Backward diagnostics: outliers and influential points’).

The R-square value for the Model is very low at 0.05, which may indicate that this is not a very good data or model. The reason for this can be value of significant variables is only from 1 to 5 representing the intensity level, and the range is very limited. Moreover, the data is from a survey and transformed into this dataset, so we may need information such as the amount of money the participant spend on entertainment, appearance instead of the level. Another possibility for this can be we don’t have enough sufficient related variables from the beginning, so in the future of improvement, we may consider creating an individual survey only focus on the finance behavior and spending about participant.

However, despite the low R-square value, these two variables are still significant for determine the finance behavior based on the likelihood estimates table as they all have very low p-value.

Prediction with Final Model

Since the final model has only two significant variables: Enter_spd and Spd_lk, the first person for prediction has 1 for Enter_spd and 2 for Spd_lk, and second person has 5 for Enter_spd and 5 for Spd_lk. The result from the final model is much better than the one from full model:

- First person: phat = 0.52759, lcl=0.44282 and ucl=0.61079
- Second person: phat=0.17965, lcl=0.13140 and ucl=0.24071

Therefore, the probability of first person is a money saver is from 55.7% to 84.2%, and the probability of second person is a money saver is from 14% to 27%.

The estimates from the regression for these predictions are:

Log(p/(1-p)) = 0.8129 – 0.2302 Enter_spd – 0.2362 Spd_lk p-value for Enter_spd is 0.0037 and for Spd_lk is 0.0026, so they are all very significant to the model.

These results make sense as one person should most likely to be a money saver if he or she doesn't spend much for the entertainment and the appearance.

Classification and Predictive performance

By adding the sensitivity with specificity together, the highest one is at prob level of 0.30 with 118, so 0.30 will be the cutoff value and the measurement for phat for the classification metric (Appendix 'Classification table').

A new binary variable pred_y is created for any rows with phat value greater than 0.30, then pred_y=1 otherwise it is 0.

Using the classification for the predictions I made previously with final model, the first phat is 0.52759 which is greater than 0.3, so pred_y=1 and this person can be identified as a money saver, but for second phat is only 0.17965 and less than 0.3 so pred_y=0, and this person is not a money saver.

The following are calculated based on the classification metric (Appendix 'Classification metric'):
Sensitivity = 0.832, Specificity = 0.635, Accuracy = 0.711
Precision = 0.589, F-metric = 0.69

The sensitivity is close to 1 which is good, but the accuracy is only 0.635 so there are still about 37% of the negatives are not classified correctly, which is only average. The Specificity is 0.711 which is ok as well but the precision is only 0.589 so about 41% in the predicted positives are false positives, and the low value of precision also result in the value of F-metric.

In conclusion, the performance result is not perfect but not bad either, so to improve in the future, the possibility ways can be changing the seed value while splitting the data, changing the percentage of how we want to split it into, requesting more relevant variables and information. Based on the low R-square value above, I think the best and significant way is to build up a more relevant dataset with wider range of value.

Conclusion

My result indicates that the consideration of one person is either money saver or spender is based on how much he or she are willing to spend on the entertainment and appearance (Enter_spd, Spd_lk) from the survey data. This outcome can be very reasonable as the participants in the survey are from 15 to 30 in age and with a lot of energy and thinking for entertainment and dressing up pretty.

The result from credit.com shows that the top 8 of a typical monthly budget for the households in U.S. from 2016 are Housing (33%), Transportation (16%), Food (13%), Pension and insurance (11%), Health care (8%), Other (7%), Entertainment (5%), Apparel and services (3%). Even though entertainment is only ranked at 7 and appearance related at 8, but others are mostly primary goods and necessities. (Sullivan). Therefore, once people are financial fulfilling for the primary goods and willing to spend money to the optional, they may be in the opposite of saving.

5. Future Work

Manish Singh section:

Huge organizations like, for example, Google, Apple, Facebook, Amazon always like to go an extra mile for their employees. They can use a model like this to analyze if their employees are money savers or not and have guidance sessions to help one's who are not and make them understand the importance of saving.
Applying this model on their dataset or even merging the two and analyzing the output would be a fun task to do.

Michal Chowaniak section:

I would like to find a similar study which would include amounts people spend, available budget for non-fixed spending categories. I would like to merge that study with our dataset and produce more detailed outcome.

Priyank Beno Cerejo section

For Future work i'll also include more variables because it was really interesting to see how data could be used to interpret such amazing results. Also, having a frequency of doing things like Shopping done once a week or once a month could also affect data. Including those variables in the future will really make the analysis more interesting.

Rushabh Shah section:

For future work, I would like to explore this same data, but with people from more than just one country, and people in a large age range. The age range of 15-30 is very restrictive and is not great at helping predict overall spending and saving habits because this age range includes mainly very young people, who have fewer financial responsibilities, as compared to adults aged 25-50, who have financial responsibilities such as a mortgage, families, children and car payments. This would help us develop a predictive model better suited to the general population, not just a small age group.

Valentine Silvester Correia section:

In future I would like to include other variables present in the original dataset with the current analysis that I produced. Also, I would try to perform imputation on the missing data for more precise analysis.

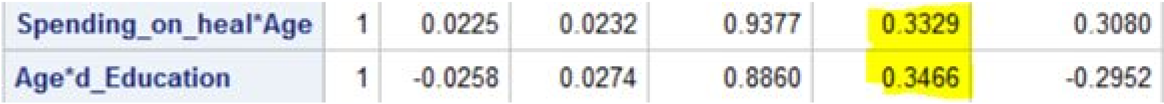
Yesheng Qin section:

The dataset I have for this model is from a very large survey with various topics, and I extracted one of them. The value of most variables is only indicated as intensity level of how strongly the participants agree on certain things instead of a specific number such as the roughly spending on entertainment. Therefore, in the future we may need to find a survey with one topic related to people spending and the survey question should target at specific amount of money that participants investment into specific field such as entertainment, shopping mall, luxuries and so on.

6. Appendix

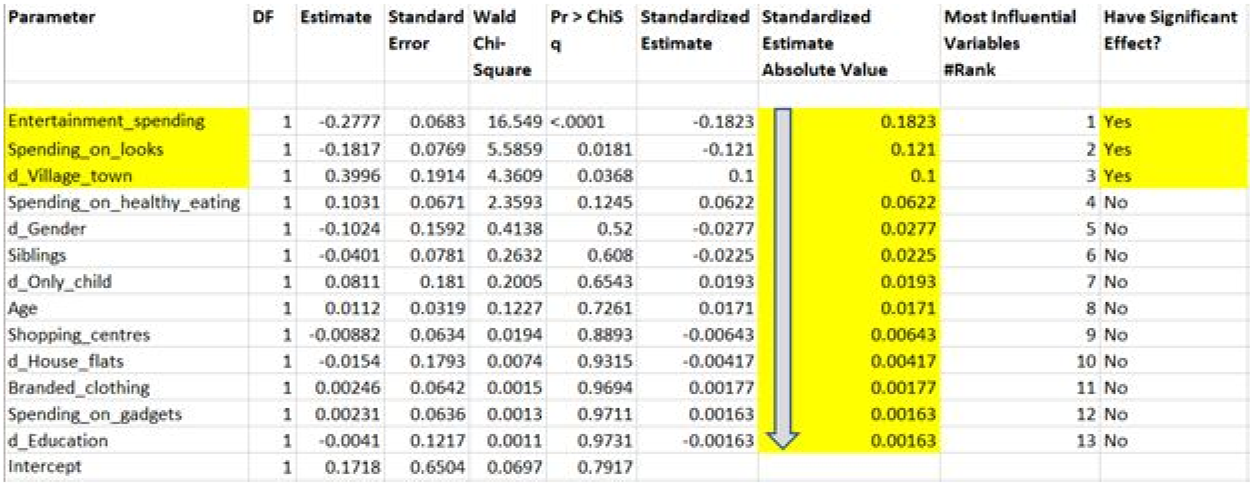
Manish Singh section:

Figure 1



Spending_on_heal*Age	1	0.0225	0.0232	0.9377	0.3329	0.3080
Age*d_Education	1	-0.0258	0.0274	0.8860	0.3466	-0.2952

Figure 2



Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Standardized Estimate Absolute Value	Most Influential Variables #Rank	Have Significant Effect?
Entertainment_spending	1	-0.2777	0.0683	16.549	<.0001	-0.1823	0.1823	1	Yes
Spending_on_looks	1	-0.1817	0.0769	5.5859	0.0181	-0.121	0.121	2	Yes
d_Village_town	1	0.3996	0.1914	4.3609	0.0368	0.1	0.1	3	Yes
Spending_on_healthy_eating	1	0.1031	0.0671	2.3593	0.1245	0.0622	0.0622	4	No
d_Gender	1	-0.1024	0.1592	0.4138	0.52	-0.0277	0.0277	5	No
Siblings	1	-0.0401	0.0781	0.2632	0.608	-0.0225	0.0225	6	No
d_Only_child	1	0.0811	0.181	0.2005	0.6543	0.0193	0.0193	7	No
Age	1	0.0112	0.0319	0.1227	0.7261	0.0171	0.0171	8	No
Shopping_centres	1	-0.00882	0.0634	0.0194	0.8893	-0.00643	0.00643	9	No
d_House_flats	1	-0.0154	0.1793	0.0074	0.9315	-0.00417	0.00417	10	No
Branded_clothing	1	0.00246	0.0642	0.0015	0.9694	0.00177	0.00177	11	No
Spending_on_gadgets	1	0.00231	0.0636	0.0013	0.9711	0.00163	0.00163	12	No
d_Education	1	-0.0041	0.1217	0.0011	0.9731	-0.00163	0.00163	13	No
Intercept	1	0.1718	0.6504	0.0697	0.7917				

Figure 3

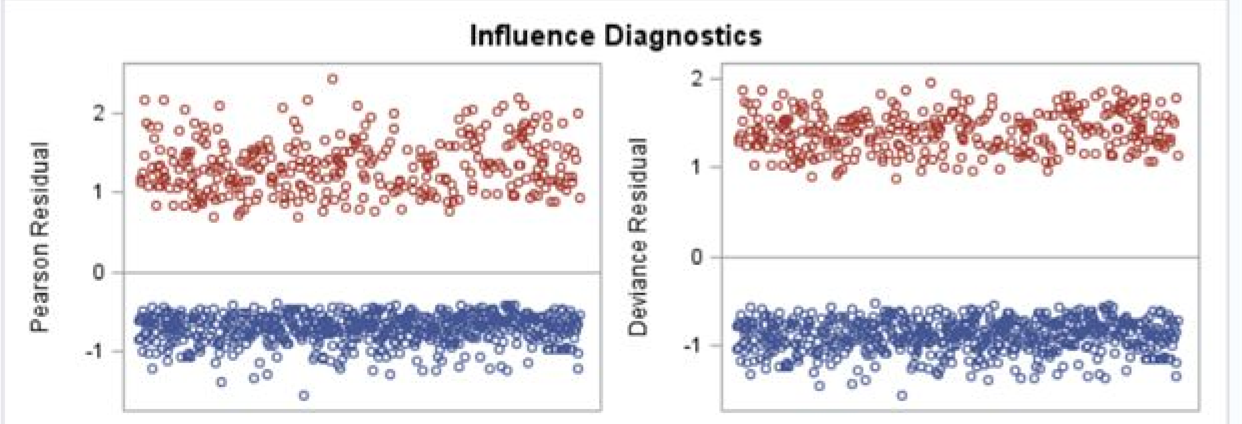
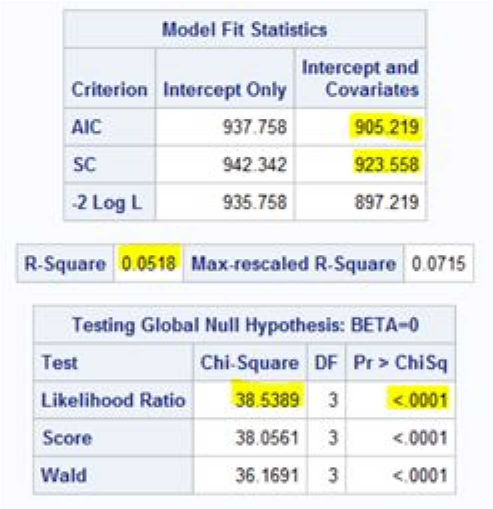


Figure 4



Model Fit Statistics

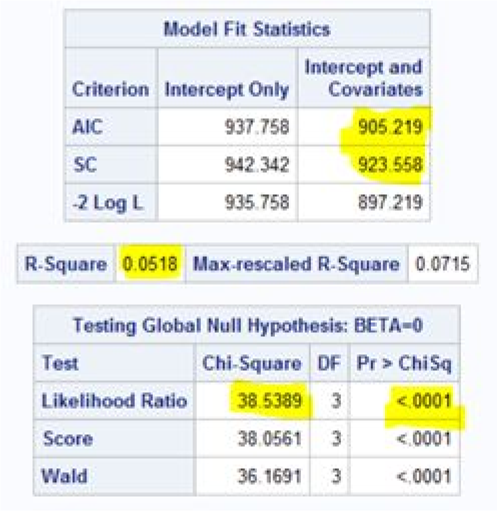
Criterion	Intercept Only	Intercept and Covariates
AIC	937.758	905.219
SC	942.342	923.558
-2 Log L	935.758	897.219

R-Square	0.0518	Max-rescaled R-Square	0.0715

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	38.5389	3	<.0001
Score	38.0561	3	<.0001
Wald	36.1691	3	<.0001

Figure 5



Model Fit Statistics

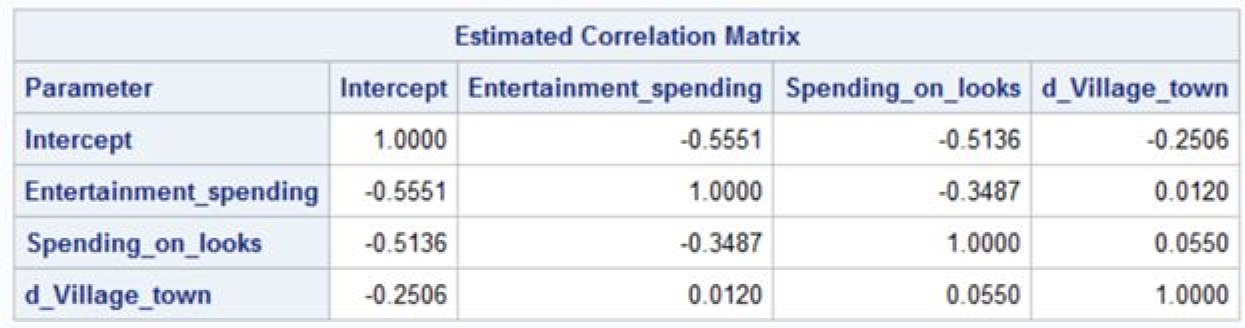
Criterion	Intercept Only	Intercept and Covariates
AIC	937.758	905.219
SC	942.342	923.558
-2 Log L	935.758	897.219

R-Square	0.0518	Max-rescaled R-Square	0.0715

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	38.5389	3	<.0001
Score	38.0561	3	<.0001
Wald	36.1691	3	<.0001

Figure 6



Estimated Correlation Matrix

Parameter	Intercept	Entertainment_spending	Spending_on_looks	d_Village_town
Intercept	1.0000	-0.5551	-0.5136	-0.2506
Entertainment_spending	-0.5551	1.0000	-0.3487	0.0120
Spending_on_looks	-0.5136	-0.3487	1.0000	0.0550
d_Village_town	-0.2506	0.0120	0.0550	1.0000

Figure 7

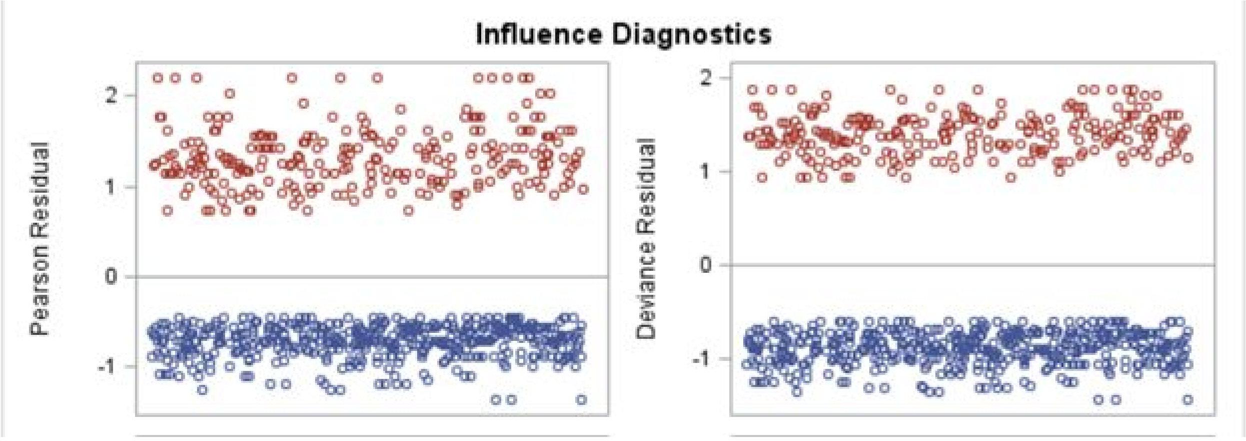
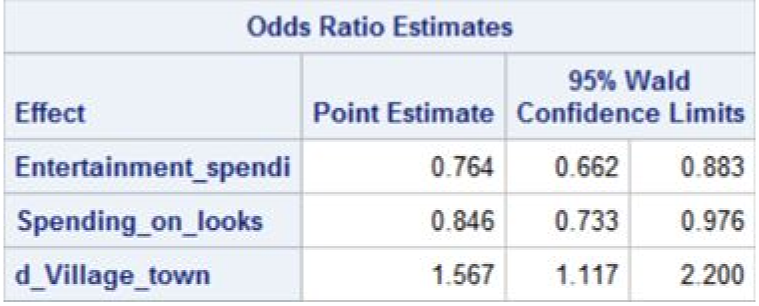
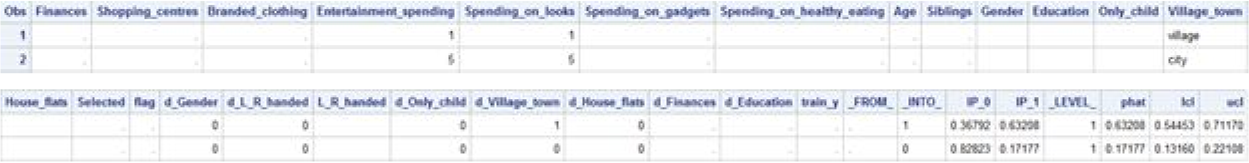


Figure 8



Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Entertainment_spendi	0.764	0.662	0.883
Spending_on_looks	0.846	0.733	0.976
d_Village_town	1.567	1.117	2.200

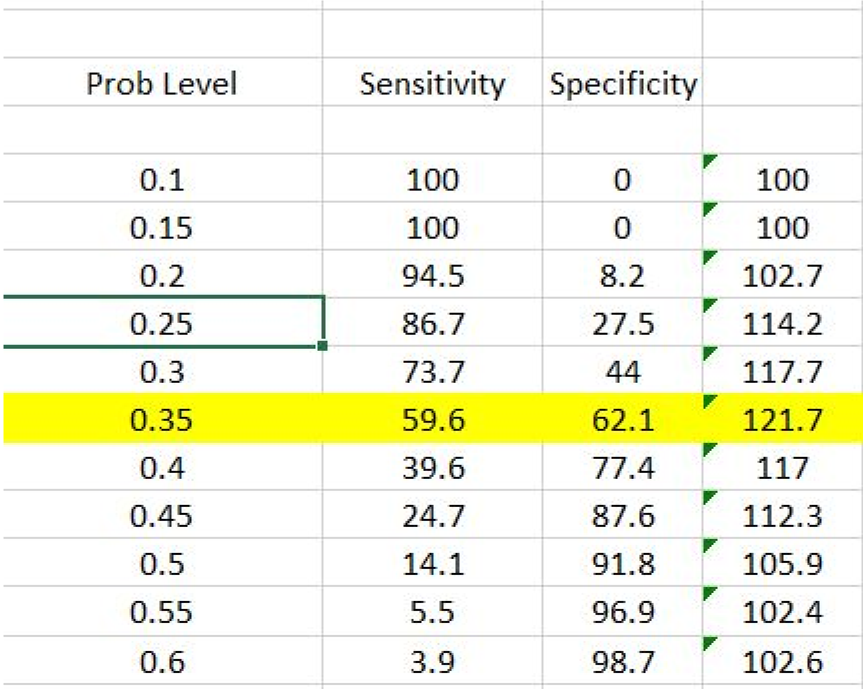
Figure 9



Obs	Finances	Shopping_centres	Branded_clothing	Entertainment_spending	Spending_on_looks	Spending_on_gadgets	Spending_on_healthy_eating	Age	Siblings	Gender	Education	Only_child	Village_town
1	.	.	.	1	1	.	.	.	.				village
2	.	.	.	5	5	.	.	.	.				city

House_flats	Selected	flag	d_Gender	d_L_R_handed	L_R_handed	d_Only_child	d_Village_town	d_House_flats	d_Finances	d_Education	train_y	_FROM_	_INTO_	IP_0	IP_1	_LEVEL_	phat	lcl	ucl
	.	.	0	0		0	1	0	.	.	.	.	1	0.36792	0.63208	1	0.63208	0.54453	0.71170
	.	.	0	0		0	0	0	.	.	.	.	0	0.82823	0.17177	1	0.17177	0.13160	0.22108

Figure 10



Prob Level	Sensitivity	Specificity	
0.1	100	0	100
0.15	100	0	100
0.2	94.5	8.2	102.7
0.25	86.7	27.5	114.2
0.3	73.7	44	117.7
0.35	59.6	62.1	121.7
0.4	39.6	77.4	117
0.45	24.7	87.6	112.3
0.5	14.1	91.8	105.9
0.55	5.5	96.9	102.4
0.6	3.9	98.7	102.6

Figure 11

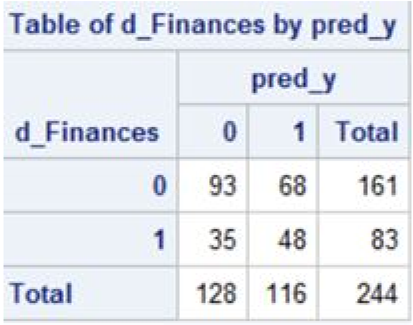


Table of d_Finances by pred_y

	pred_y		
d_Finances	0	1	Total
0	93	68	161
1	35	48	83
Total	128	116	244

Figure 12

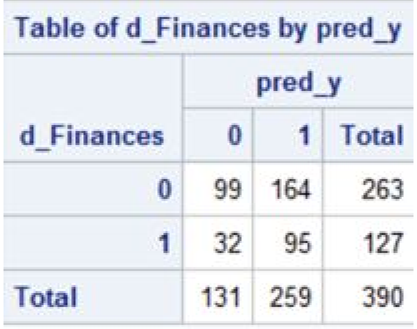
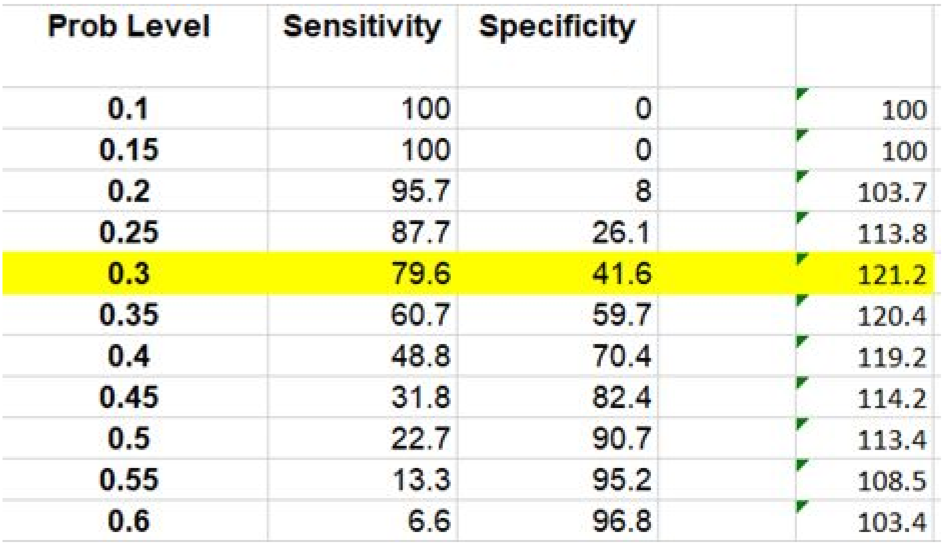


Table of d_Finances by pred_y

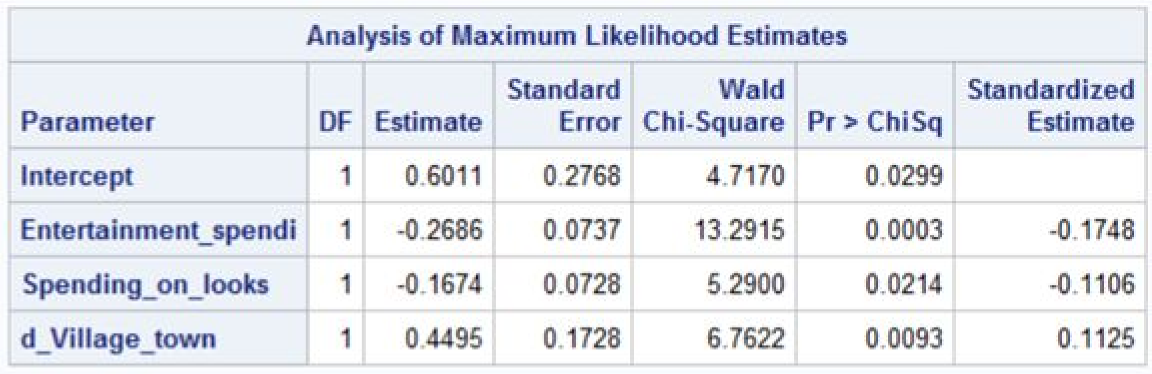
	pred_y		
d_Finances	0	1	Total
0	99	164	263
1	32	95	127
Total	131	259	390

Figure 13



Prob Level	Sensitivity	Specificity		
0.1	100	0		100
0.15	100	0		100
0.2	95.7	8		103.7
0.25	87.7	26.1		113.8
0.3	79.6	41.6		121.2
0.35	60.7	59.7		120.4
0.4	48.8	70.4		119.2
0.45	31.8	82.4		114.2
0.5	22.7	90.7		113.4
0.55	13.3	95.2		108.5
0.6	6.6	96.8		103.4

Figure 14



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	0.6011	0.2768	4.7170	0.0299	
Entertainment_spendi	1	-0.2686	0.0737	13.2915	0.0003	-0.1748
Spending_on_looks	1	-0.1674	0.0728	5.2900	0.0214	-0.1106
d_Village_town	1	0.4495	0.1728	6.7622	0.0093	0.1125

Michal Chowaniak section:

Figure 1. Finances before and after binning

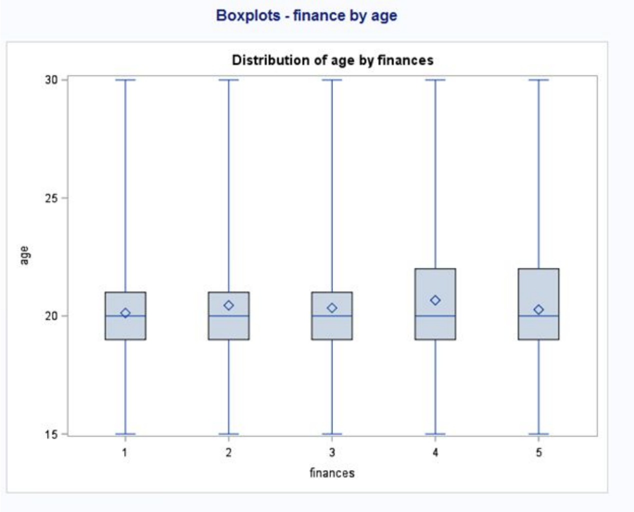


Figure 2. Finances Binned

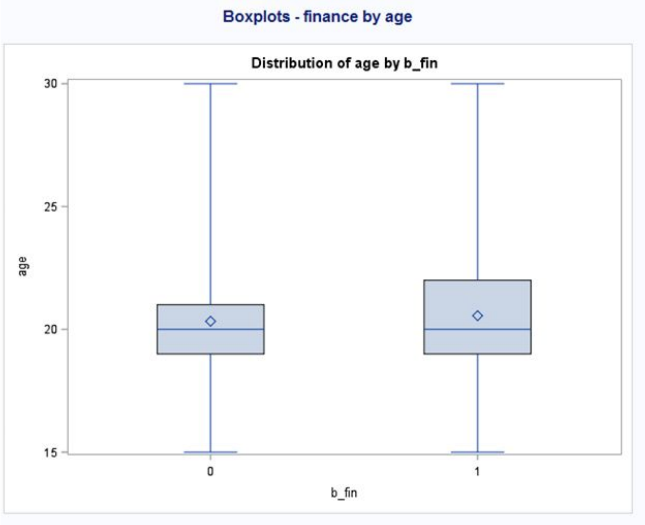
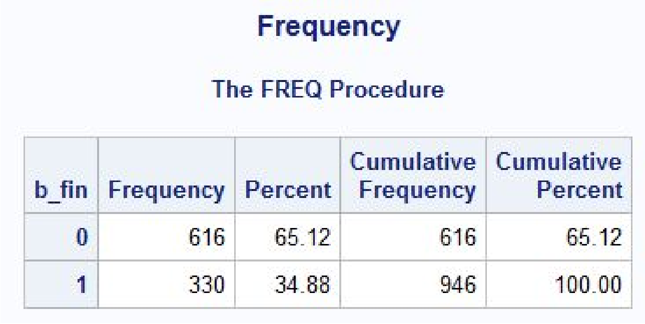


Figure 3. Frequency

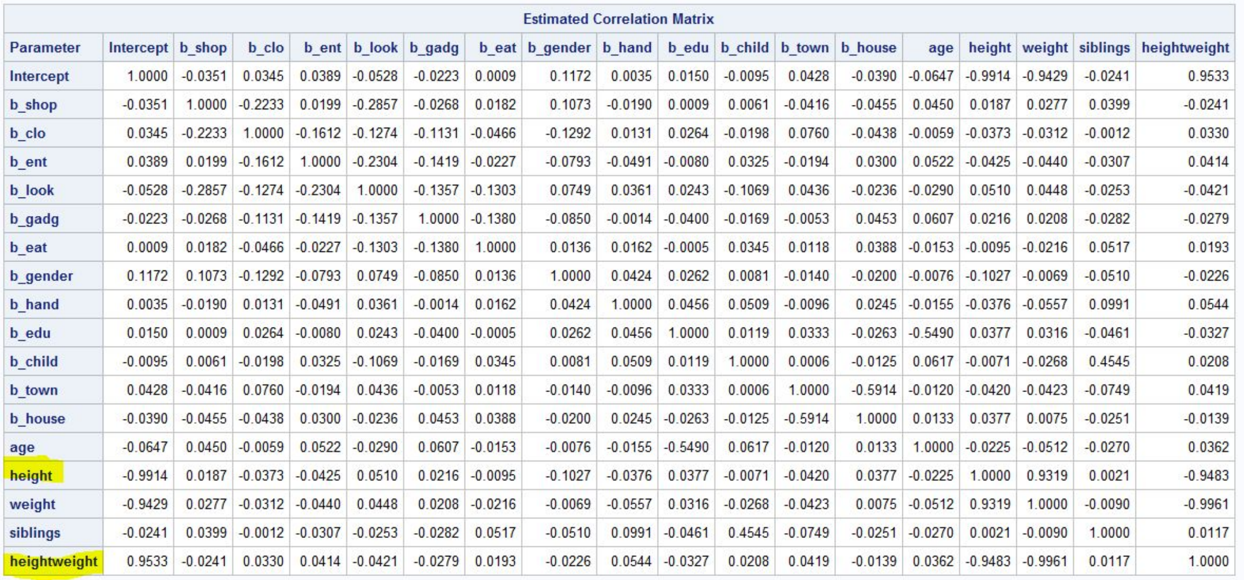


Frequency

The FREQ Procedure

b_fin	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	616	65.12	616	65.12
1	330	34.88	946	100.00

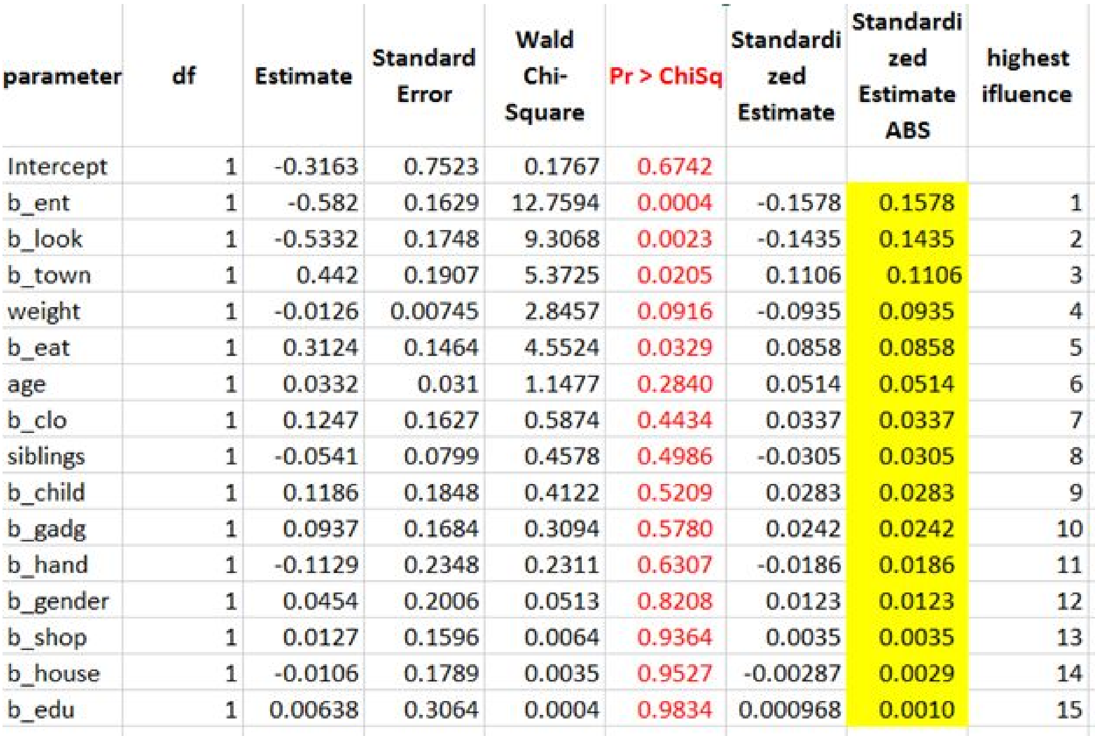
Figure 4. Correlation



Estimated Correlation Matrix

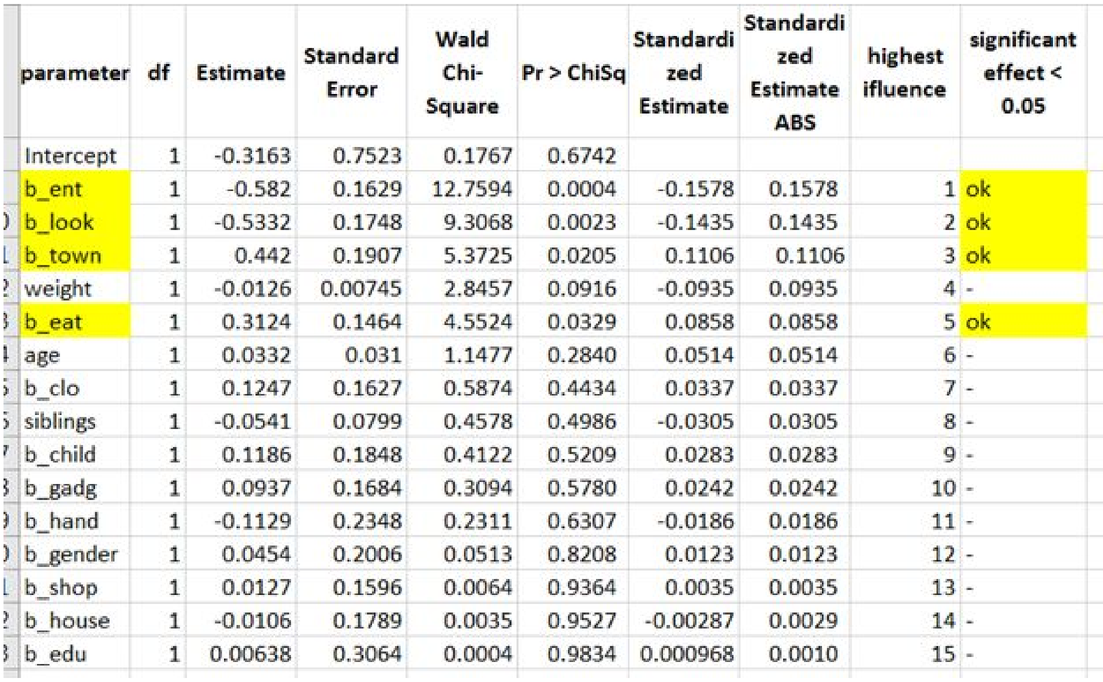
Parameter	Intercept	b_shop	b_clo	b_ent	b_look	b_gadg	b_eat	b_gender	b_hand	b_edu	b_child	b_town	b_house	age	height	weight	siblings	heightweight
Intercept	1.0000	-0.0351	0.0345	0.0389	-0.0528	-0.0223	0.0009	0.1172	0.0035	0.0150	-0.0095	0.0428	-0.0390	-0.0647	-0.9914	-0.9429	-0.0241	0.9533
b_shop	-0.0351	1.0000	-0.2233	0.0199	-0.2857	-0.0268	0.0182	0.1073	-0.0190	0.0009	0.0061	-0.0416	-0.0455	0.0450	0.0187	0.0277	0.0399	-0.0241
b_clo	0.0345	-0.2233	1.0000	-0.1612	-0.1274	-0.1131	-0.0466	-0.1292	0.0131	0.0264	-0.0198	0.0760	-0.0438	-0.0059	-0.0373	-0.0312	-0.0012	0.0330
b_ent	0.0389	0.0199	-0.1612	1.0000	-0.2304	-0.1419	-0.0227	-0.0793	-0.0491	-0.0080	0.0325	-0.0194	0.0300	0.0522	-0.0425	-0.0440	-0.0307	0.0414
b_look	-0.0528	-0.2857	-0.1274	-0.2304	1.0000	-0.1357	-0.1303	0.0749	0.0361	0.0243	-0.1069	0.0436	-0.0236	-0.0290	0.0510	0.0448	-0.0253	-0.0421
b_gadg	-0.0223	-0.0268	-0.1131	-0.1419	-0.1357	1.0000	-0.1380	-0.0850	-0.0014	-0.0400	-0.0169	-0.0053	0.0453	0.0607	0.0216	0.0208	-0.0282	-0.0279
b_eat	0.0009	0.0182	-0.0466	-0.0227	-0.1303	-0.1380	1.0000	0.0136	0.0162	-0.0005	0.0345	0.0118	0.0388	-0.0153	-0.0095	-0.0216	0.0517	0.0193
b_gender	0.1172	0.1073	-0.1292	-0.0793	0.0749	-0.0850	0.0136	1.0000	0.0424	0.0262	0.0081	-0.0140	-0.0200	-0.0076	-0.1027	-0.0069	-0.0510	-0.0226
b_hand	0.0035	-0.0190	0.0131	-0.0491	0.0361	-0.0014	0.0162	0.0424	1.0000	0.0456	0.0509	-0.0096	0.0245	-0.0155	-0.0376	-0.0557	0.0991	0.0544
b_edu	0.0150	0.0009	0.0264	-0.0080	0.0243	-0.0400	-0.0005	0.0262	0.0456	1.0000	0.0119	0.0333	-0.0263	-0.5490	0.0377	0.0316	-0.0461	-0.0327
b_child	-0.0095	0.0061	-0.0198	0.0325	-0.1069	-0.0169	0.0345	0.0081	0.0509	0.0119	1.0000	0.0006	-0.0125	0.0617	-0.0071	-0.0268	0.4545	0.0208
b_town	0.0428	-0.0416	0.0760	-0.0194	0.0436	-0.0053	0.0118	-0.0140	-0.0096	0.0333	0.0006	1.0000	-0.5914	-0.0120	-0.0420	-0.0423	-0.0749	0.0419
b_house	-0.0390	-0.0455	-0.0438	0.0300	-0.0236	0.0453	0.0388	-0.0200	0.0245	-0.0263	-0.0125	-0.5914	1.0000	0.0133	0.0377	0.0075	-0.0251	-0.0139
age	-0.0647	0.0450	-0.0059	0.0522	-0.0290	0.0607	-0.0153	-0.0076	-0.0155	-0.5490	0.0617	-0.0120	0.0133	1.0000	-0.0225	-0.0512	-0.0270	0.0362
height	-0.9914	0.0187	-0.0373	-0.0425	0.0510	0.0216	-0.0095	-0.1027	-0.0376	0.0377	-0.0071	-0.0420	0.0377	-0.0225	1.0000	0.9319	0.0021	-0.9483
weight	-0.9429	0.0277	-0.0312	-0.0440	0.0448	0.0208	-0.0216	-0.0069	-0.0557	0.0316	-0.0268	-0.0423	0.0075	-0.0512	0.9319	1.0000	-0.0090	-0.9961
siblings	-0.0241	0.0399	-0.0012	-0.0307	-0.0253	-0.0282	0.0517	-0.0510	0.0991	-0.0461	0.4545	-0.0749	-0.0251	-0.0270	0.0021	-0.0090	1.0000	0.0117
heightweight	0.9533	-0.0241	0.0330	0.0414	-0.0421	-0.0279	0.0193	-0.0226	0.0544	-0.0327	0.0208	0.0419	-0.0139	0.0362	-0.9483	-0.9961	0.0117	1.0000

Figure 5. Highest influence variables



parameter	df	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Standardized Estimate ABS	highest ifluence
Intercept	1	-0.3163	0.7523	0.1767	0.6742			
b_ent	1	-0.582	0.1629	12.7594	0.0004	-0.1578	0.1578	1
b_look	1	-0.5332	0.1748	9.3068	0.0023	-0.1435	0.1435	2
b_town	1	0.442	0.1907	5.3725	0.0205	0.1106	0.1106	3
weight	1	-0.0126	0.00745	2.8457	0.0916	-0.0935	0.0935	4
b_eat	1	0.3124	0.1464	4.5524	0.0329	0.0858	0.0858	5
age	1	0.0332	0.031	1.1477	0.2840	0.0514	0.0514	6
b_clo	1	0.1247	0.1627	0.5874	0.4434	0.0337	0.0337	7
siblings	1	-0.0541	0.0799	0.4578	0.4986	-0.0305	0.0305	8
b_child	1	0.1186	0.1848	0.4122	0.5209	0.0283	0.0283	9
b_gadg	1	0.0937	0.1684	0.3094	0.5780	0.0242	0.0242	10
b_hand	1	-0.1129	0.2348	0.2311	0.6307	-0.0186	0.0186	11
b_gender	1	0.0454	0.2006	0.0513	0.8208	0.0123	0.0123	12
b_shop	1	0.0127	0.1596	0.0064	0.9364	0.0035	0.0035	13
b_house	1	-0.0106	0.1789	0.0035	0.9527	-0.00287	0.0029	14
b_edu	1	0.00638	0.3064	0.0004	0.9834	0.000968	0.0010	15

Figure 6. Significant Predictors



parameter	df	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Standardized Estimate ABS	highest ifluence	significant effect < 0.05
Intercept	1	-0.3163	0.7523	0.1767	0.6742				
b_ent	1	-0.582	0.1629	12.7594	0.0004	-0.1578	0.1578	1	ok
b_look	1	-0.5332	0.1748	9.3068	0.0023	-0.1435	0.1435	2	ok
b_town	1	0.442	0.1907	5.3725	0.0205	0.1106	0.1106	3	ok
weight	1	-0.0126	0.00745	2.8457	0.0916	-0.0935	0.0935	4	-
b_eat	1	0.3124	0.1464	4.5524	0.0329	0.0858	0.0858	5	ok
age	1	0.0332	0.031	1.1477	0.2840	0.0514	0.0514	6	-
b_clo	1	0.1247	0.1627	0.5874	0.4434	0.0337	0.0337	7	-
siblings	1	-0.0541	0.0799	0.4578	0.4986	-0.0305	0.0305	8	-
b_child	1	0.1186	0.1848	0.4122	0.5209	0.0283	0.0283	9	-
b_gadg	1	0.0937	0.1684	0.3094	0.5780	0.0242	0.0242	10	-
b_hand	1	-0.1129	0.2348	0.2311	0.6307	-0.0186	0.0186	11	-
b_gender	1	0.0454	0.2006	0.0513	0.8208	0.0123	0.0123	12	-
b_shop	1	0.0127	0.1596	0.0064	0.9364	0.0035	0.0035	13	-
b_house	1	-0.0106	0.1789	0.0035	0.9527	-0.00287	0.0029	14	-
b_edu	1	0.00638	0.3064	0.0004	0.9834	0.000968	0.0010	15	-

Figure 7. Outliers

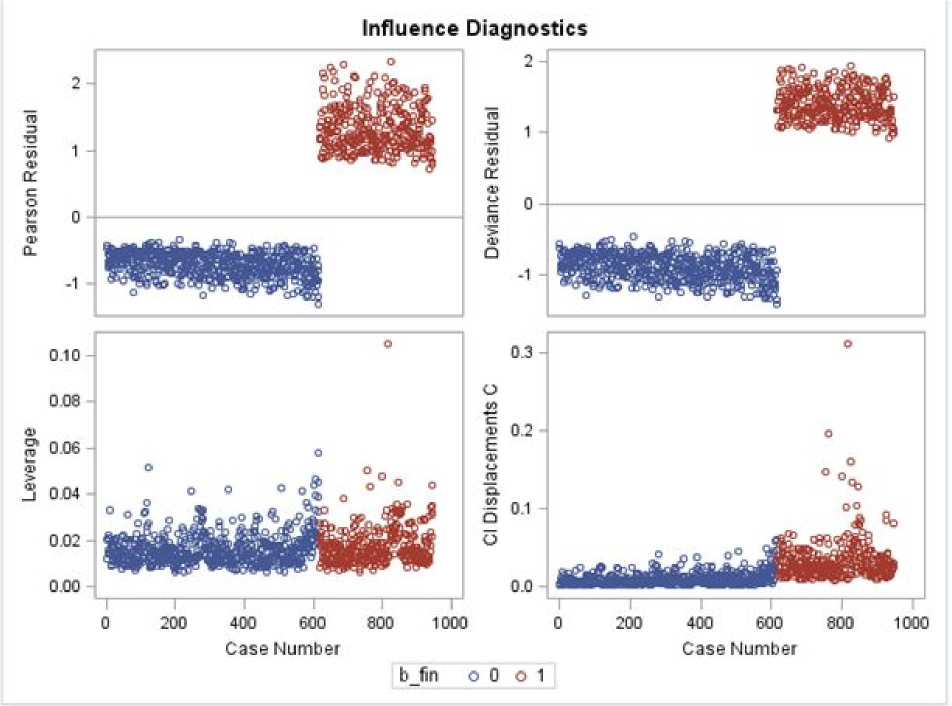


Figure 8. Influential Points

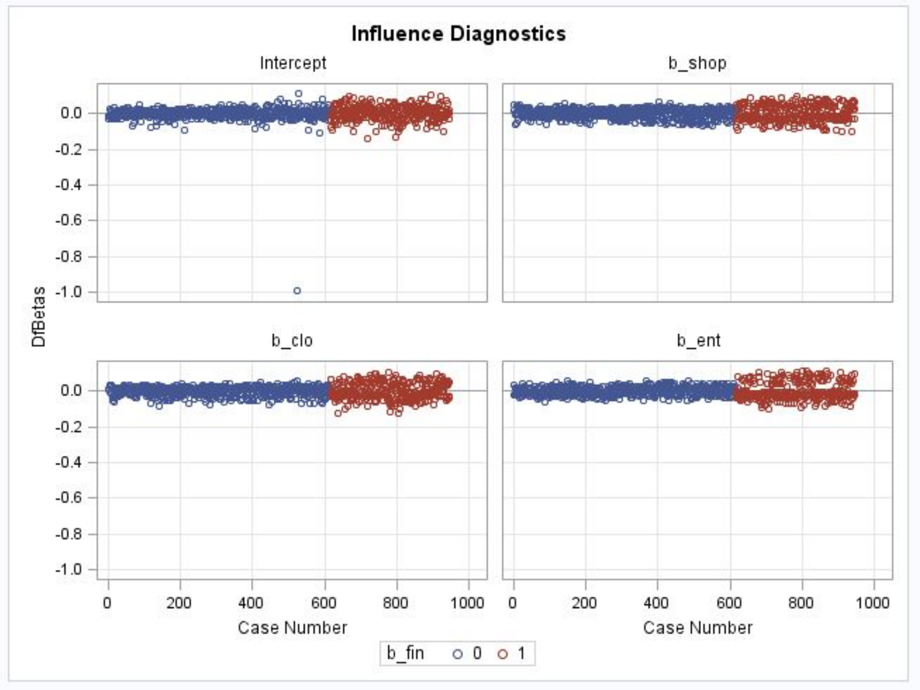


Figure 9. Influential Points

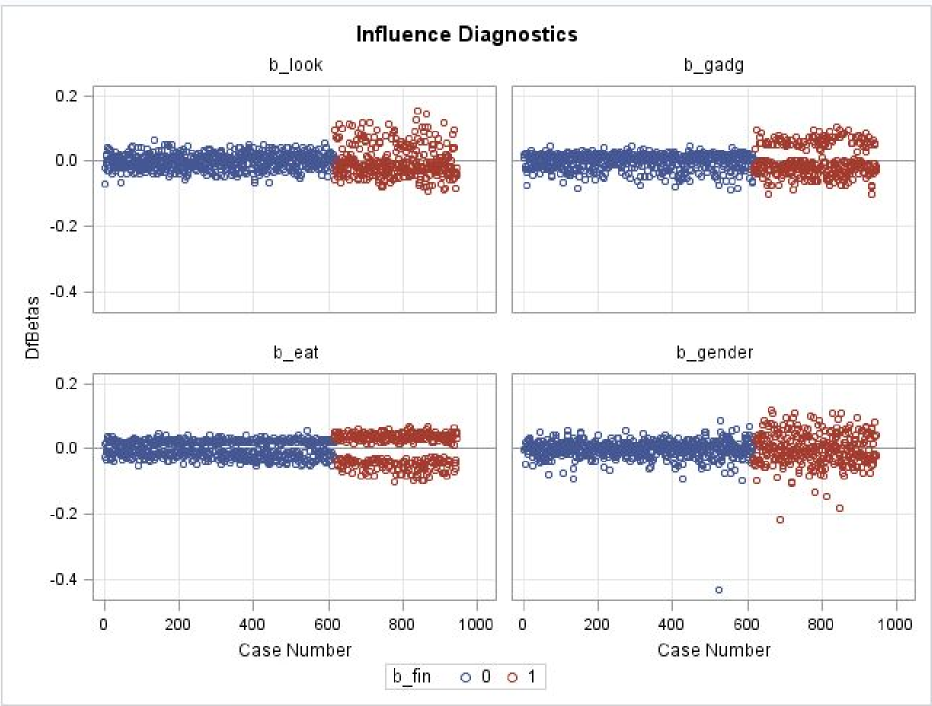


Figure 10. Influential Points

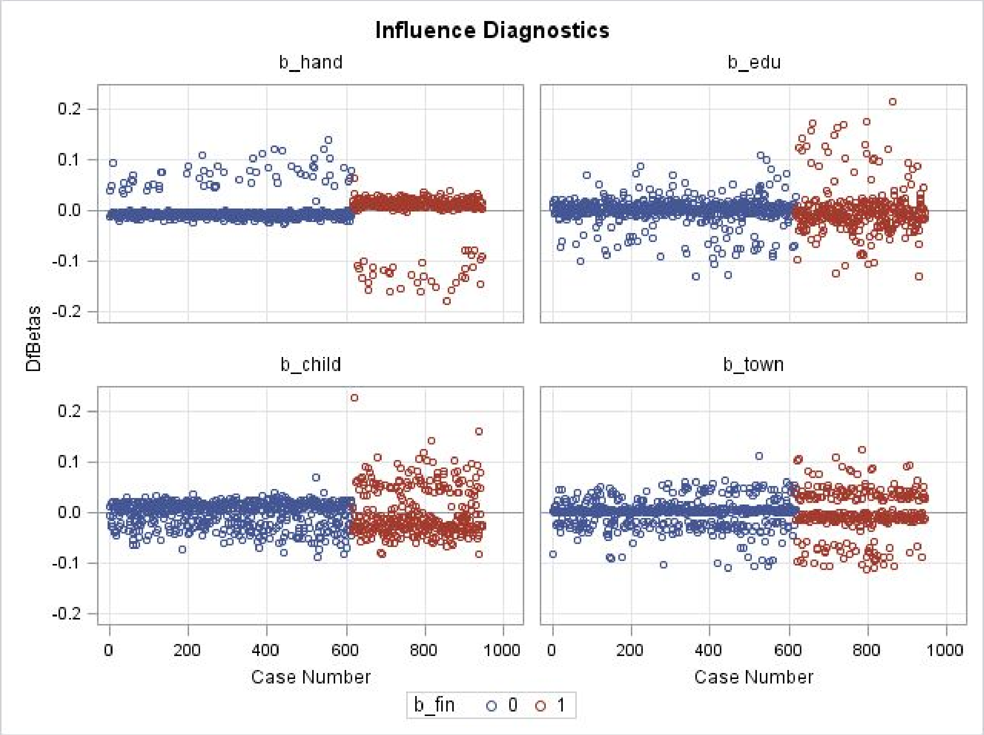


Figure 11. Influential Points

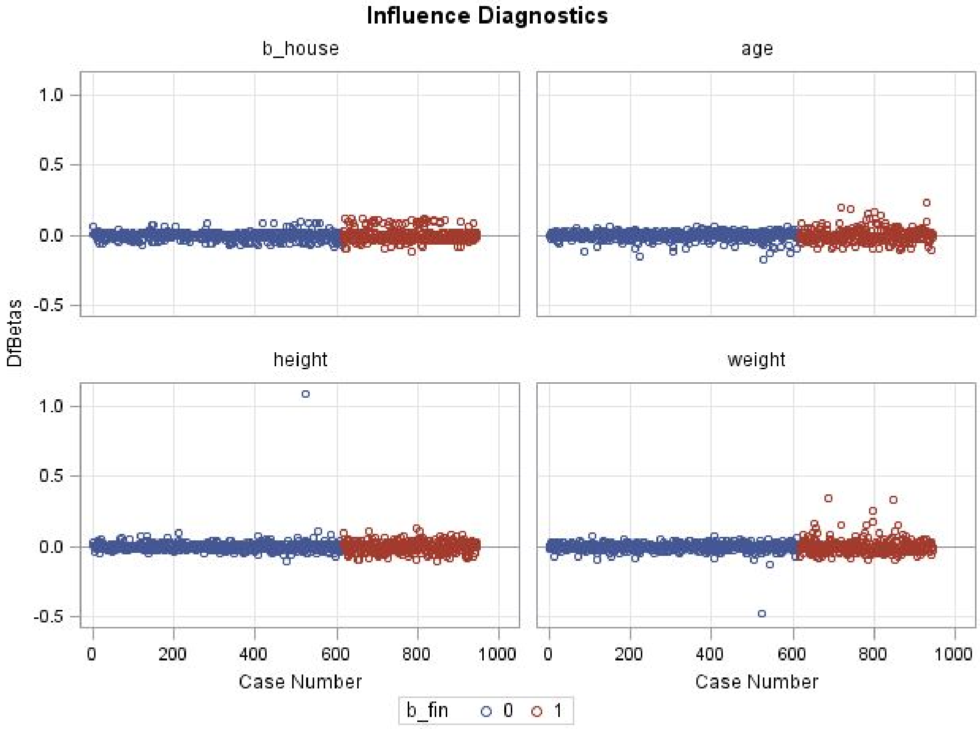


Figure 12. Influential Points

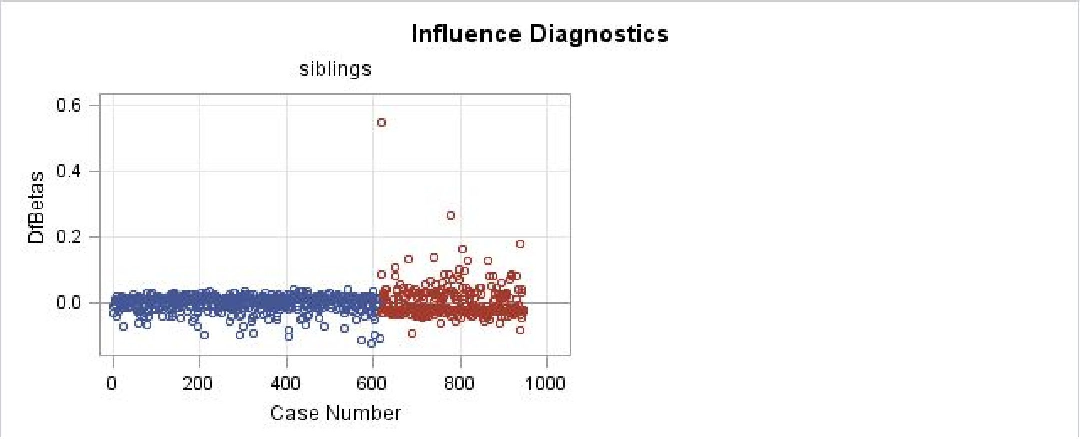
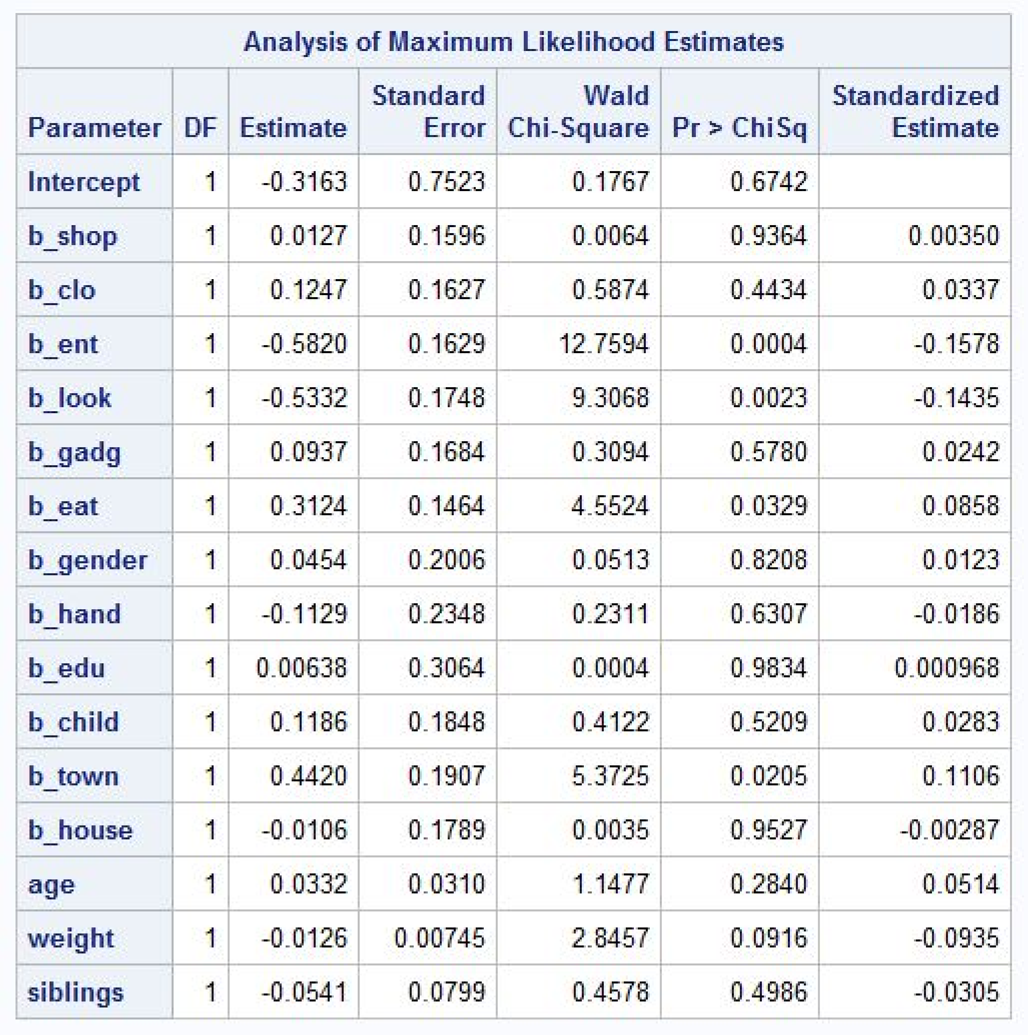


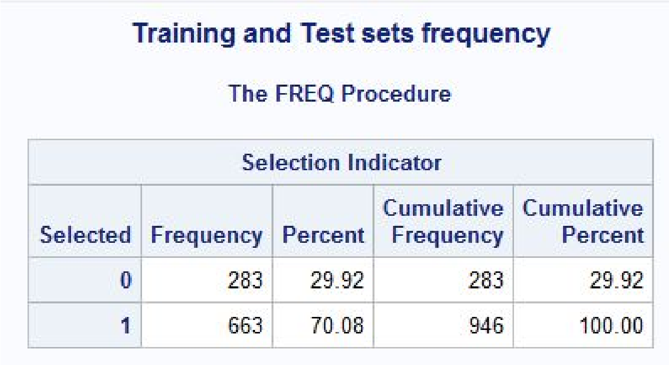
Figure 13. Full Model



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-0.3163	0.7523	0.1767	0.6742	
b_shop	1	0.0127	0.1596	0.0064	0.9364	0.00350
b_clo	1	0.1247	0.1627	0.5874	0.4434	0.0337
b_ent	1	-0.5820	0.1629	12.7594	0.0004	-0.1578
b_look	1	-0.5332	0.1748	9.3068	0.0023	-0.1435
b_gadg	1	0.0937	0.1684	0.3094	0.5780	0.0242
b_eat	1	0.3124	0.1464	4.5524	0.0329	0.0858
b_gender	1	0.0454	0.2006	0.0513	0.8208	0.0123
b_hand	1	-0.1129	0.2348	0.2311	0.6307	-0.0186
b_edu	1	0.00638	0.3064	0.0004	0.9834	0.000968
b_child	1	0.1186	0.1848	0.4122	0.5209	0.0283
b_town	1	0.4420	0.1907	5.3725	0.0205	0.1106
b_house	1	-0.0106	0.1789	0.0035	0.9527	-0.00287
age	1	0.0332	0.0310	1.1477	0.2840	0.0514
weight	1	-0.0126	0.00745	2.8457	0.0916	-0.0935
siblings	1	-0.0541	0.0799	0.4578	0.4986	-0.0305

Figure 14. Model B frequency

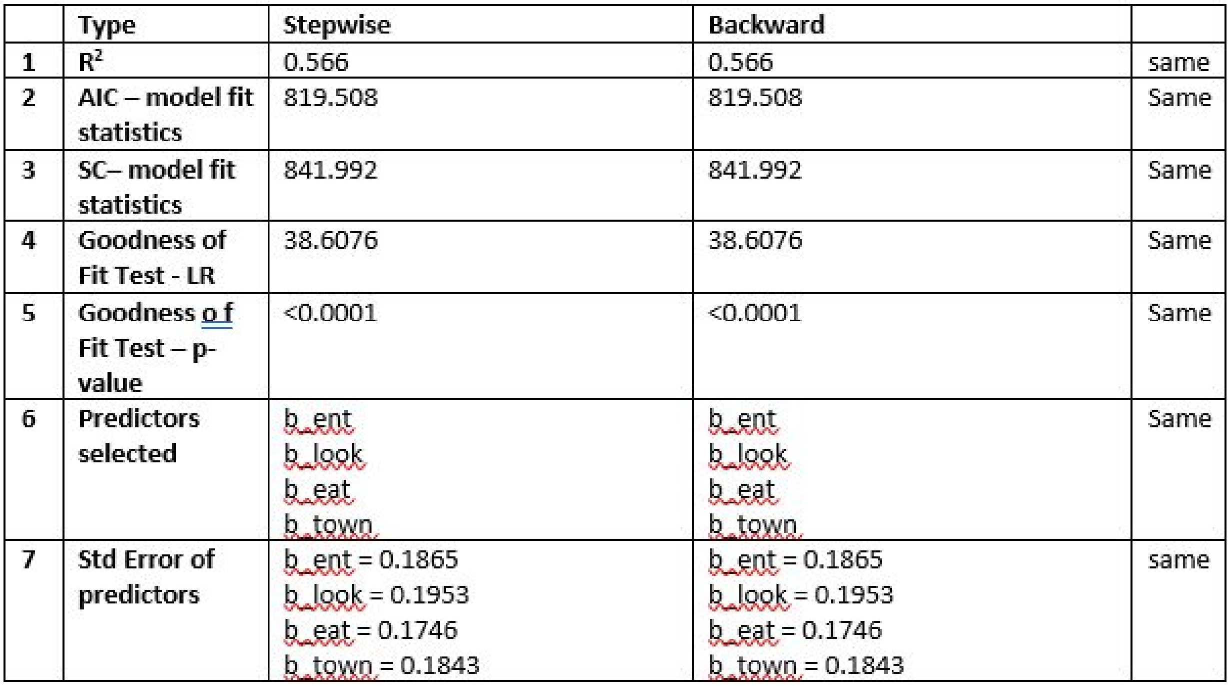


Training and Test sets frequency

The FREQ Procedure

Selection Indicator				
Selected	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	283	29.92	283	29.92
1	663	70.08	946	100.00

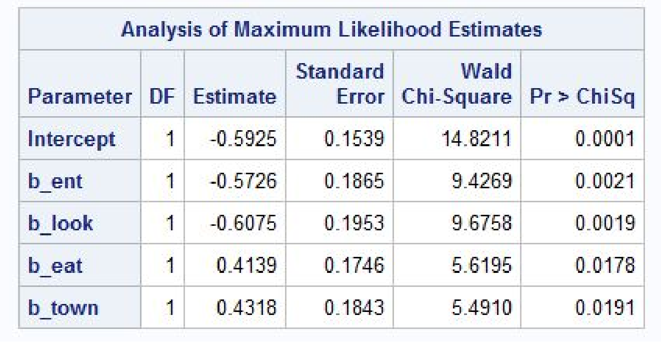
Figure 15.



	Type	Stepwise	Backward	
1	R^2	0.566	0.566	same
2	AIC – model fit statistics	819.508	819.508	Same
3	SC– model fit statistics	841.992	841.992	Same
4	Goodness of Fit Test - LR	38.6076	38.6076	Same
5	Goodness o f Fit Test – p-value	<0.0001	<0.0001	Same
6	Predictors selected	b_ent b_look b_eat b_town	b_ent b_look b_eat b_town	Same

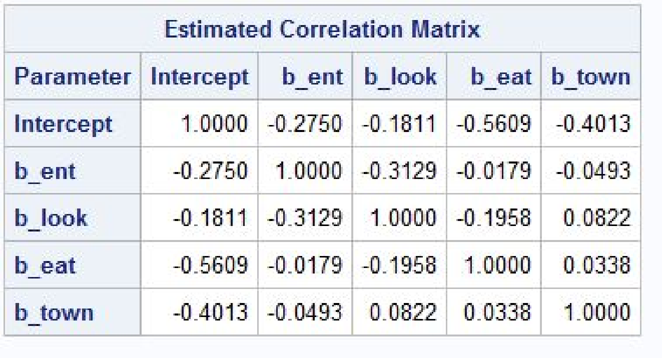
7	Std Error of predictors	b_ent = 0.1865
		b_look = 0.1953
		b_eat = 0.1746
		b_town = 0.1843
			b_ent = 0.1865
			b_look = 0.1953
			b_eat = 0.1746
			b_town = 0.1843
				same

Figure 16. Final Model



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.5925	0.1539	14.8211	0.0001
b_ent	1	-0.5726	0.1865	9.4269	0.0021
b_look	1	-0.6075	0.1953	9.6758	0.0019
b_eat	1	0.4139	0.1746	5.6195	0.0178
b_town	1	0.4318	0.1843	5.4910	0.0191

Figure 17. Final Model Correlation



Estimated Correlation Matrix					
Parameter	Intercept	b_ent	b_look	b_eat	b_town
Intercept	1.0000	-0.2750	-0.1811	-0.5609	-0.4013
b_ent	-0.2750	1.0000	-0.3129	-0.0179	-0.0493
b_look	-0.1811	-0.3129	1.0000	-0.1958	0.0822
b_eat	-0.5609	-0.0179	-0.1958	1.0000	0.0338
b_town	-0.4013	-0.0493	0.0822	0.0338	1.0000

Figure 18. Outliers

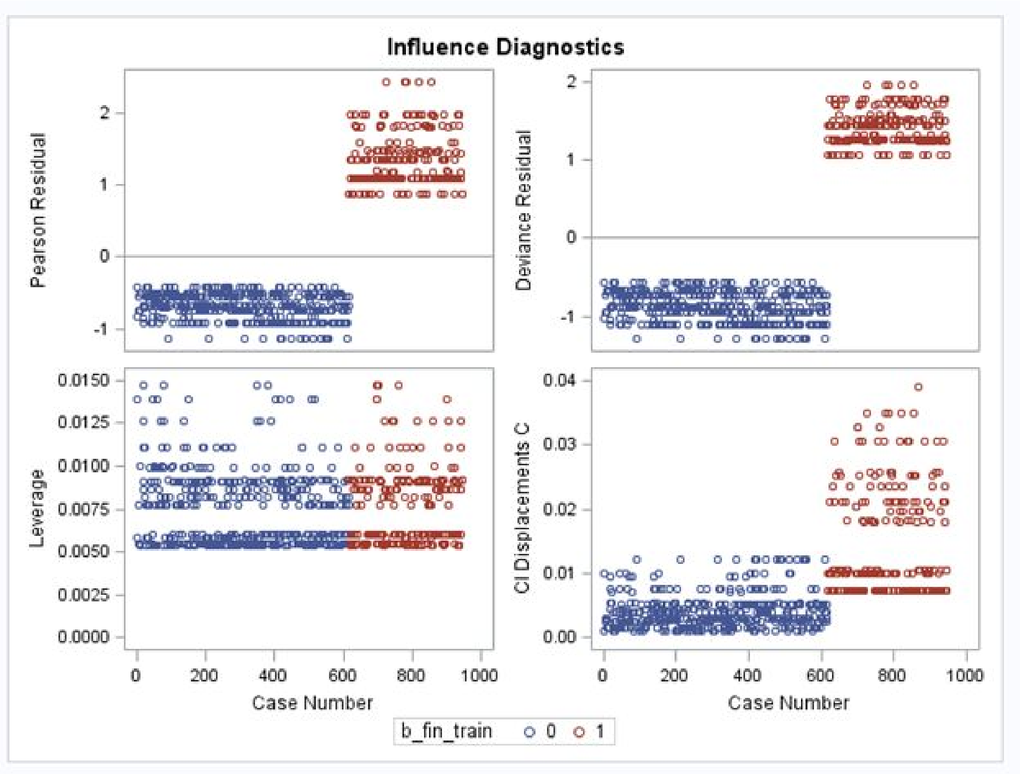
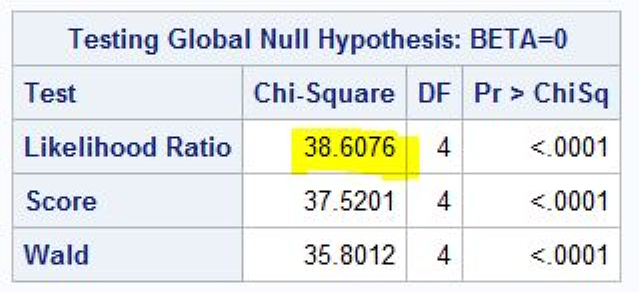
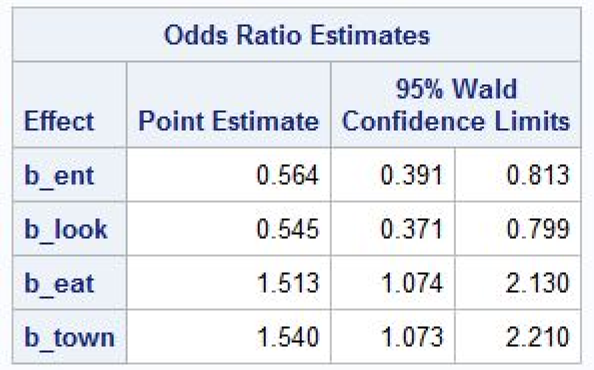


Figure 19. Global Null Hypothesis



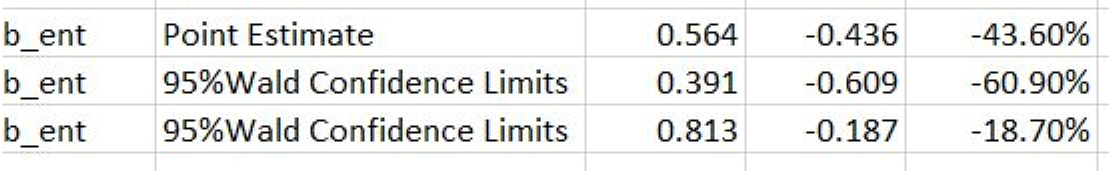
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	38.6076	4	<.0001
Score	37.5201	4	<.0001
Wald	35.8012	4	<.0001

Figure 20. Odds Ratios



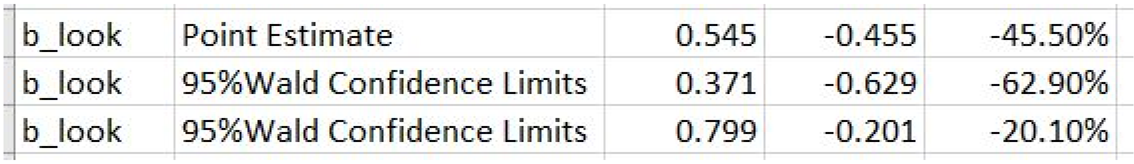
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
b_ent	0.564	0.391	0.813
b_look	0.545	0.371	0.799
b_eat	1.513	1.074	2.130
b_town	1.540	1.073	2.210

Figure 21. Odds for b_ent



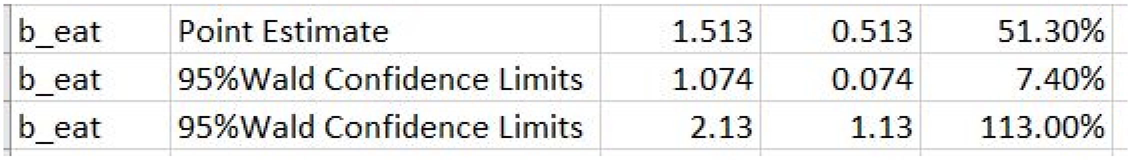
b_ent	Point Estimate	0.564	-0.436	-43.60%
b_ent	95%Wald Confidence Limits	0.391	-0.609	-60.90%
b_ent	95%Wald Confidence Limits	0.813	-0.187	-18.70%

Figure 22. Odds for b_look



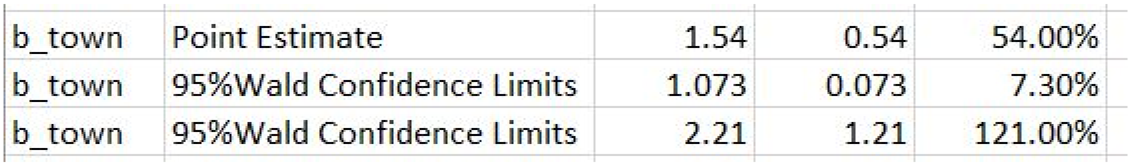
b_look	Point Estimate	0.545	-0.455	-45.50%
b_look	95%Wald Confidence Limits	0.371	-0.629	-62.90%
b_look	95%Wald Confidence Limits	0.799	-0.201	-20.10%

Figure 23. Odds for b_eat



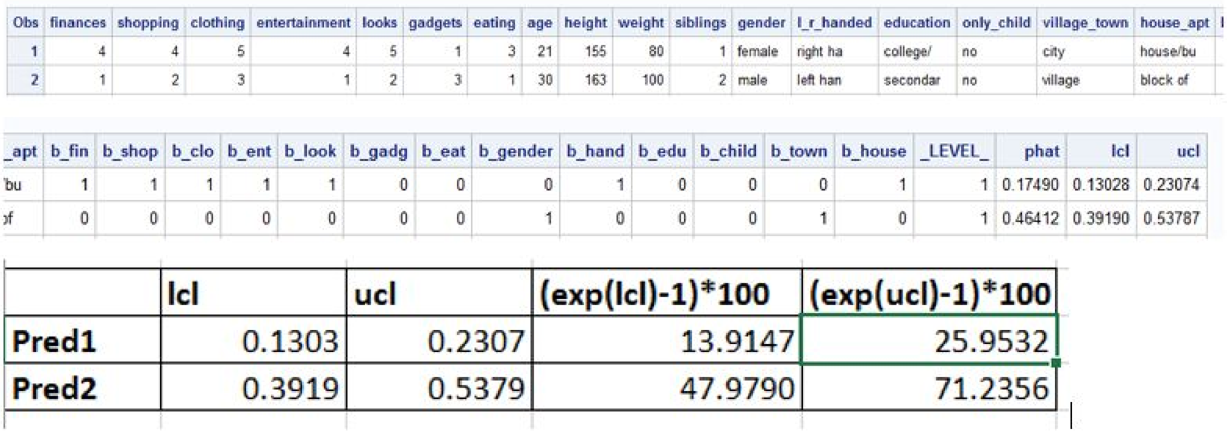
b_eat	Point Estimate	1.513	0.513	51.30%
b_eat	95%Wald Confidence Limits	1.074	0.074	7.40%
b_eat	95%Wald Confidence Limits	2.13	1.13	113.00%

Figure 24. Odds for b_town



b_town	Point Estimate	1.54	0.54	54.00%
b_town	95%Wald Confidence Limits	1.073	0.073	7.30%
b_town	95%Wald Confidence Limits	2.21	1.21	121.00%

Figure 25. Prediction Computations

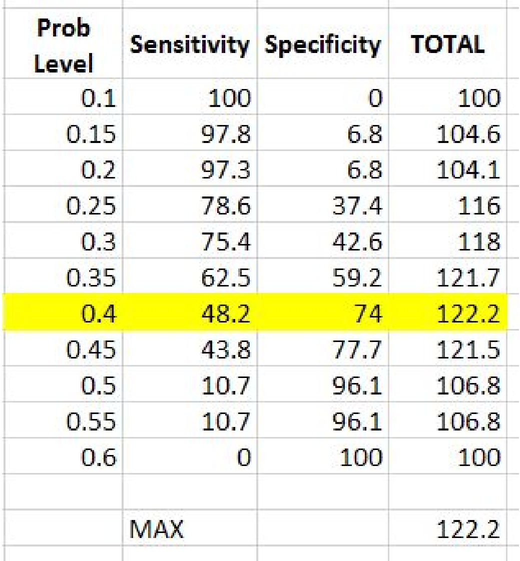


Obs	finances	shopping	clothing	entertainment	looks	gadgets	eating	age	height	weight	siblings	gender	l_r_handed	education	only_child	village_town	house_apt	l
1	4	4	5	4	5	1	3	21	155	80	1	female	right ha	college/	no	city	house/bu	
2	1	2	3	1	2	3	1	30	163	100	2	male	left han	secondar	no	village	block of	

_apt	b_fin	b_shop	b_clo	b_ent	b_look	b_gadg	b_eat	b_gender	b_hand	b_edu	b_child	b_town	b_house	_LEVEL_	phat	lcl	ucl
'bu	1	1	1	1	1	0	0	0	1	0	0	0	1	1	0.17490	0.13028	0.23074
of	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0.46412	0.39190	0.53787

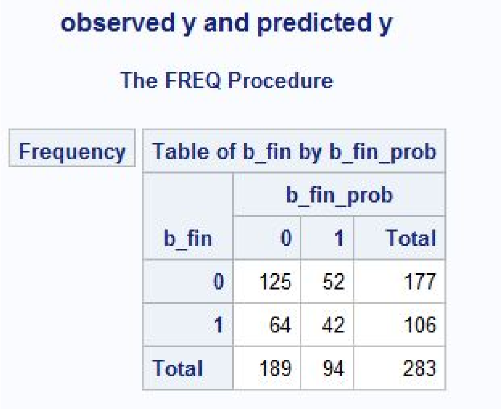
	lcl	ucl	(exp(lcl)-1)*100	(exp(ucl)-1)*100
Pred1	0.1303	0.2307	13.9147	25.9532
Pred2	0.3919	0.5379	47.9790	71.2356

Figure 26. Classification table, cut off value.



Prob Level	Sensitivity	Specificity	TOTAL
0.1	100	0	100
0.15	97.8	6.8	104.6
0.2	97.3	6.8	104.1
0.25	78.6	37.4	116
0.3	75.4	42.6	118
0.35	62.5	59.2	121.7
0.4	48.2	74	122.2
0.45	43.8	77.7	121.5
0.5	10.7	96.1	106.8
0.55	10.7	96.1	106.8
0.6	0	100	100
	MAX		122.2

Figure 27. Observed vs predicted y variable

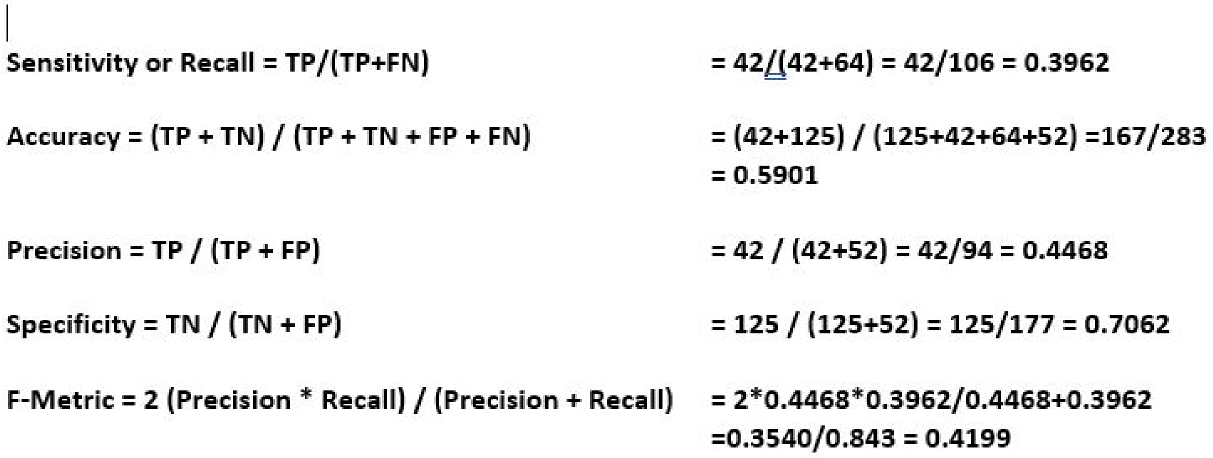


observed y and predicted y

The FREQ Procedure

Frequency	Table of b_fin by b_fin_prob			
		b_fin_prob		
	b_fin	0	1	Total
	0	125	52	177
	1	64	42	106
	Total	189	94	283

Figure 28. Measures



Sensitivity or Recall = TP/(TP+FN) **= 42/(42+64) = 42/106 = 0.3962**

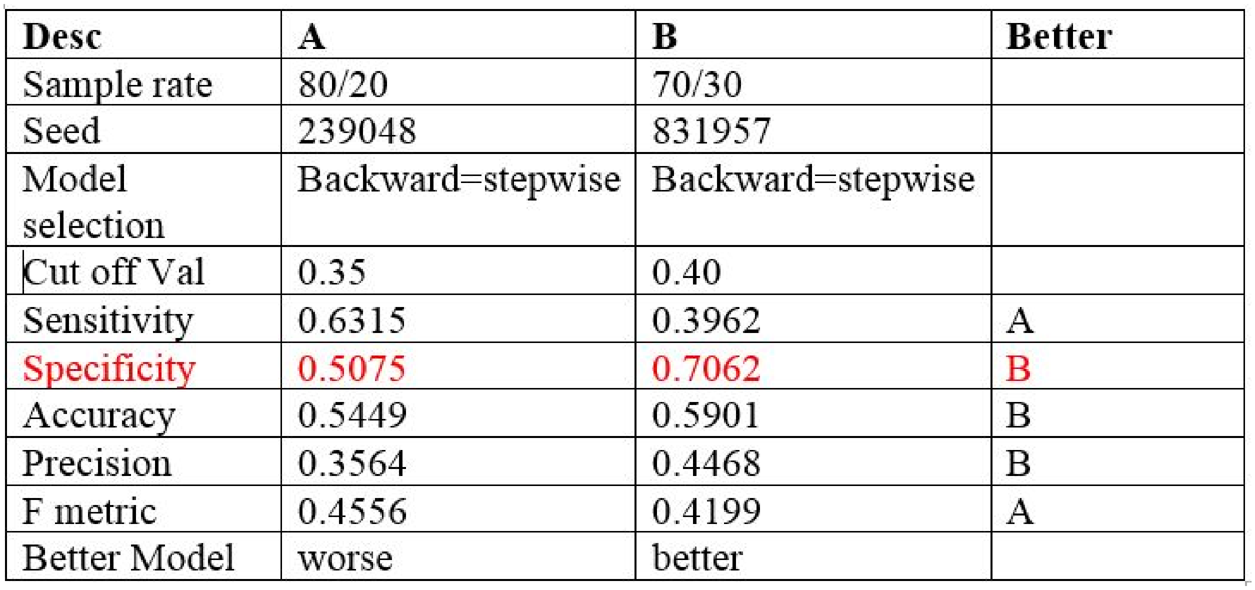
Accuracy = (TP + TN) / (TP + TN + FP + FN) **= (42+125) / (125+42+64+52) =167/283 = 0.5901**

Precision = TP / (TP + FP) **= 42 / (42+52) = 42/94 = 0.4468**

Specificity = TN / (TN + FP) **= 125 / (125+52) = 125/177 = 0.7062**

F-Metric = 2 (Precision * Recall) / (Precision + Recall) **= 2*0.4468*0.3962/0.4468+0.3962 =0.3540/0.843 = 0.4199**

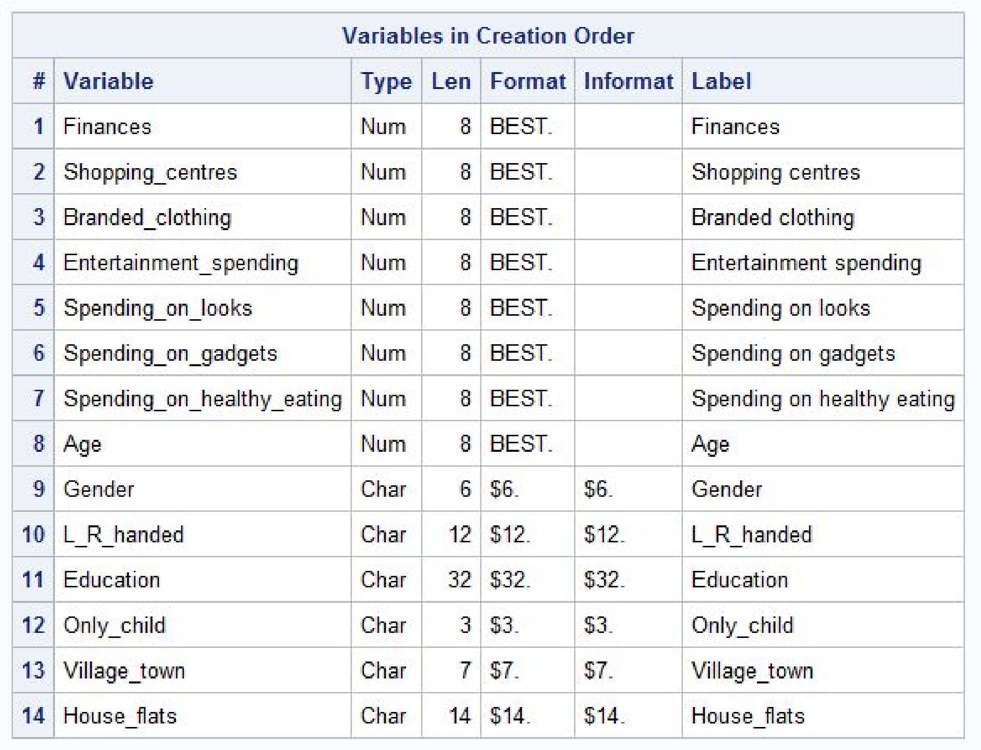
Figure 29. Model comparison



Desc	A	B	Better
Sample rate	80/20	70/30	
Seed	239048	831957	
Model selection	Backward=stepwise	Backward=stepwise	
Cut off Val	0.35	0.40	
Sensitivity	0.6315	0.3962	A
Specificity	0.5075	0.7062	B
Accuracy	0.5449	0.5901	B
Precision	0.3564	0.4468	B
F metric	0.4556	0.4199	A
Better Model	worse	better	

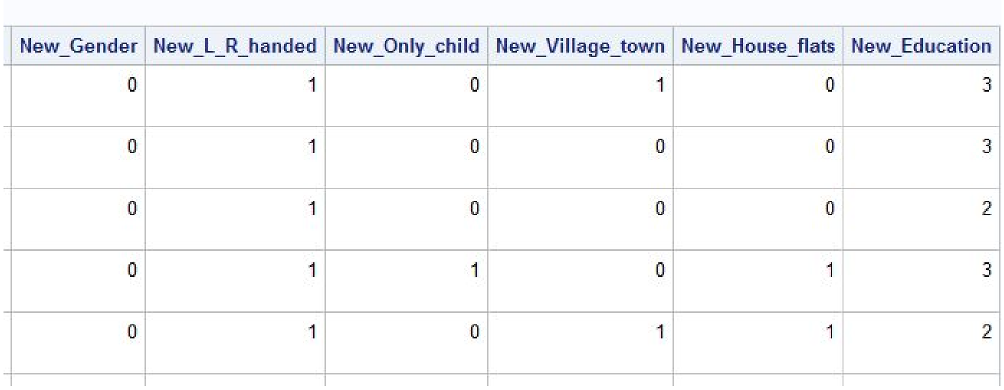
Priyank Beno Cerejo section:

Fig 1.



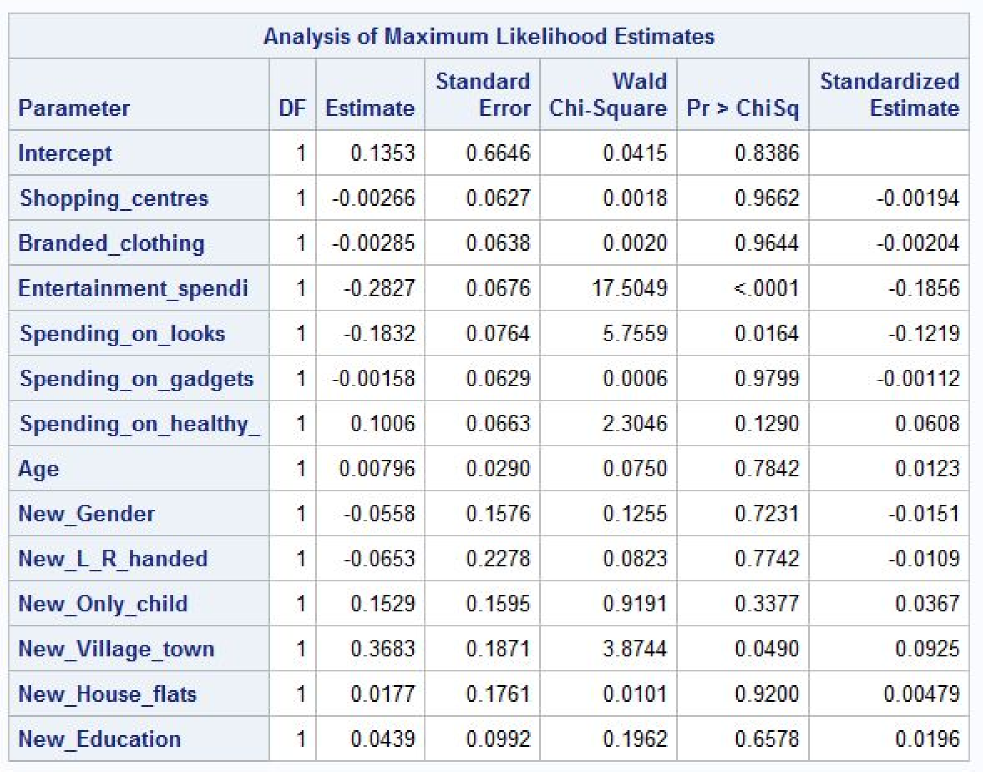
Variables in Creation Order						
#	Variable	Type	Len	Format	Informat	Label
1	Finances	Num	8	BEST.		Finances
2	Shopping_centres	Num	8	BEST.		Shopping centres
3	Branded_clothing	Num	8	BEST.		Branded clothing
4	Entertainment_spending	Num	8	BEST.		Entertainment spending
5	Spending_on_looks	Num	8	BEST.		Spending on looks
6	Spending_on_gadgets	Num	8	BEST.		Spending on gadgets
7	Spending_on_healthy_eating	Num	8	BEST.		Spending on healthy eating
8	Age	Num	8	BEST.		Age
9	Gender	Char	6	$6.	$6.	Gender
10	L_R_handed	Char	12	$12.	$12.	L_R_handed
11	Education	Char	32	$32.	$32.	Education
12	Only_child	Char	3	$3.	$3.	Only_child
13	Village_town	Char	7	$7.	$7.	Village_town
14	House_flats	Char	14	$14.	$14.	House_flats

Fig 2.



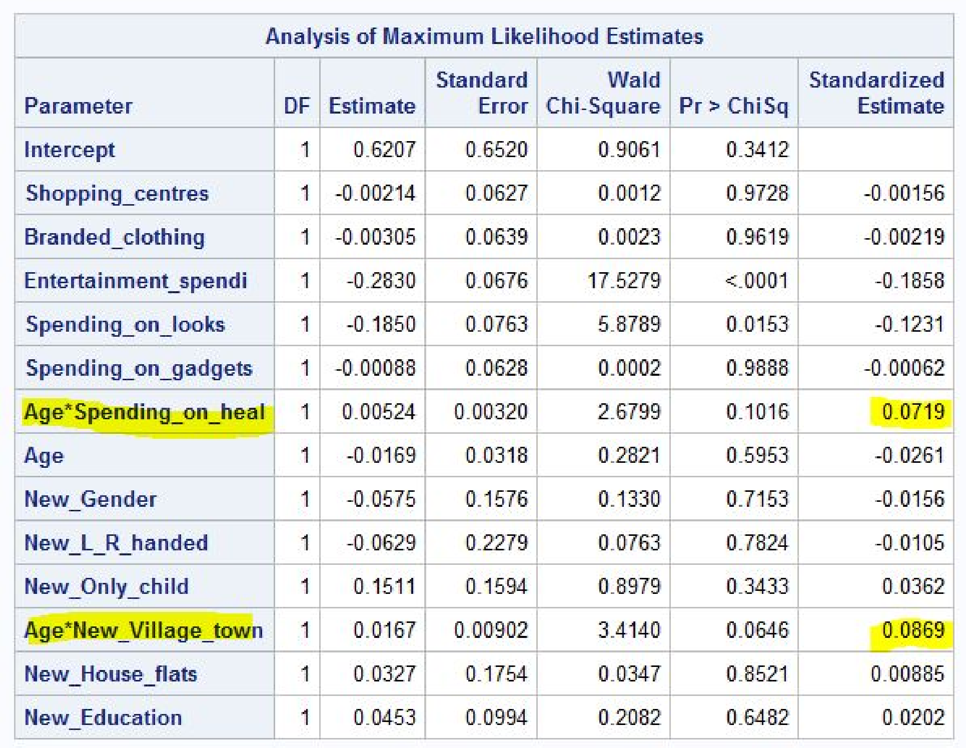
New_Gender	New_L_R_handed	New_Only_child	New_Village_town	New_House_flats	New_Education
0	1	0	1	0	3
0	1	0	0	0	3
0	1	0	0	0	2
0	1	1	0	1	3
0	1	0	1	1	2

Fig 3.



Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	0.1353	0.6646	0.0415	0.8386	
Shopping_centres	1	-0.00266	0.0627	0.0018	0.9662	-0.00194
Branded_clothing	1	-0.00285	0.0638	0.0020	0.9644	-0.00204
Entertainment_spendi	1	-0.2827	0.0676	17.5049	<.0001	-0.1856
Spending_on_looks	1	-0.1832	0.0764	5.7559	0.0164	-0.1219
Spending_on_gadgets	1	-0.00158	0.0629	0.0006	0.9799	-0.00112
Spending_on_healthy_	1	0.1006	0.0663	2.3046	0.1290	0.0608
Age	1	0.00796	0.0290	0.0750	0.7842	0.0123
New_Gender	1	-0.0558	0.1576	0.1255	0.7231	-0.0151
New_L_R_handed	1	-0.0653	0.2278	0.0823	0.7742	-0.0109
New_Only_child	1	0.1529	0.1595	0.9191	0.3377	0.0367
New_Village_town	1	0.3683	0.1871	3.8744	0.0490	0.0925
New_House_flats	1	0.0177	0.1761	0.0101	0.9200	0.00479
New_Education	1	0.0439	0.0992	0.1962	0.6578	0.0196

Fig 4.



Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	0.6207	0.6520	0.9061	0.3412	
Shopping_centres	1	-0.00214	0.0627	0.0012	0.9728	-0.00156
Branded_clothing	1	-0.00305	0.0639	0.0023	0.9619	-0.00219
Entertainment_spendi	1	-0.2830	0.0676	17.5279	<.0001	-0.1858
Spending_on_looks	1	-0.1850	0.0763	5.8789	0.0153	-0.1231
Spending_on_gadgets	1	-0.00088	0.0628	0.0002	0.9888	-0.00062
Age*Spending_on_heal	1	0.00524	0.00320	2.6799	0.1016	0.0719
Age	1	-0.0169	0.0318	0.2821	0.5953	-0.0261
New_Gender	1	-0.0575	0.1576	0.1330	0.7153	-0.0156
New_L_R_handed	1	-0.0629	0.2279	0.0763	0.7824	-0.0105
New_Only_child	1	0.1511	0.1594	0.8979	0.3433	0.0362
Age*New_Village_town	1	0.0167	0.00902	3.4140	0.0646	0.0869
New_House_flats	1	0.0327	0.1754	0.0347	0.8521	0.00885
New_Education	1	0.0453	0.0994	0.2082	0.6482	0.0202

Fig 5.

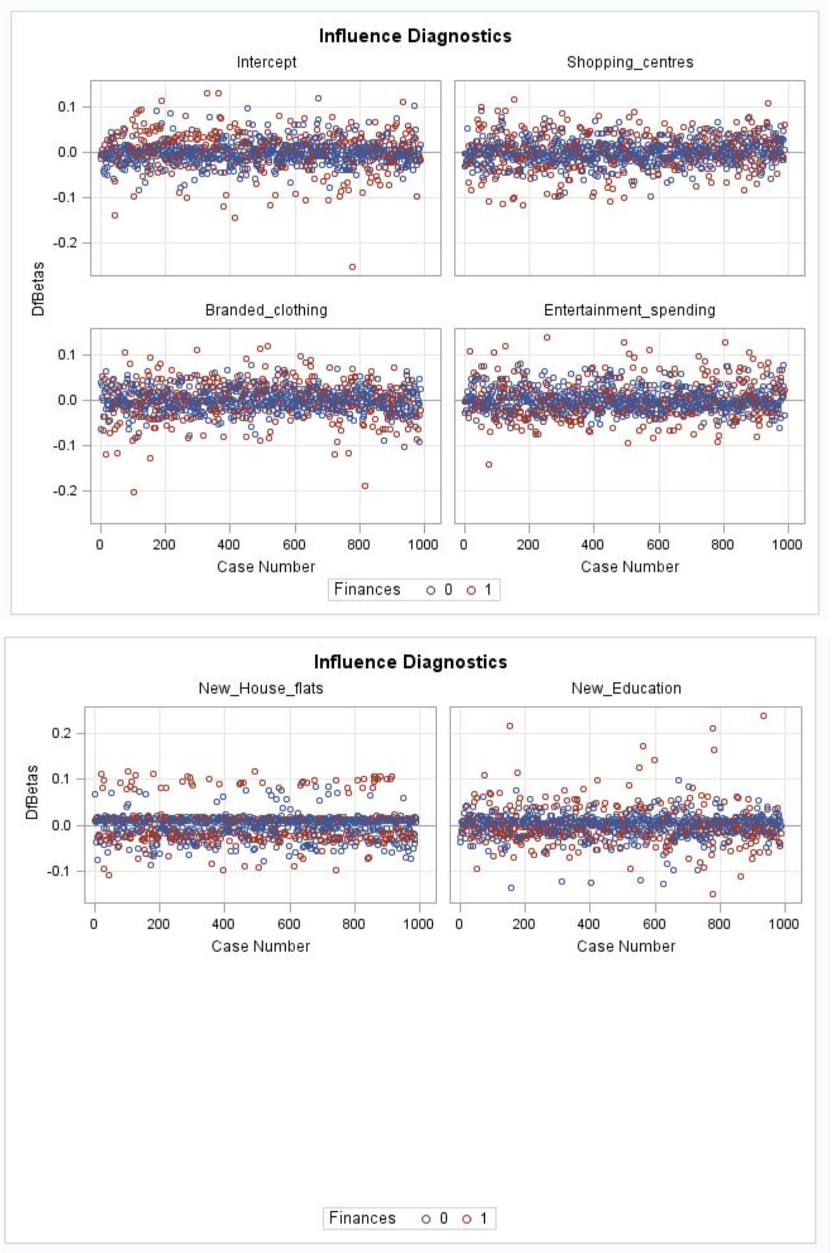
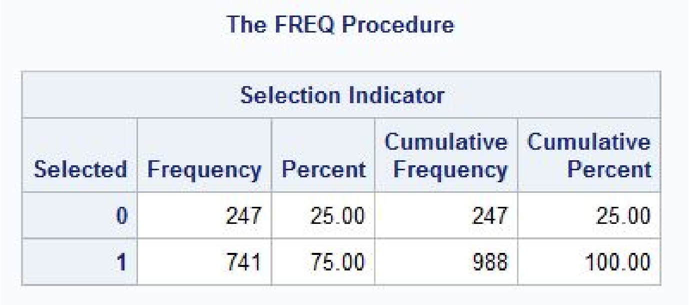


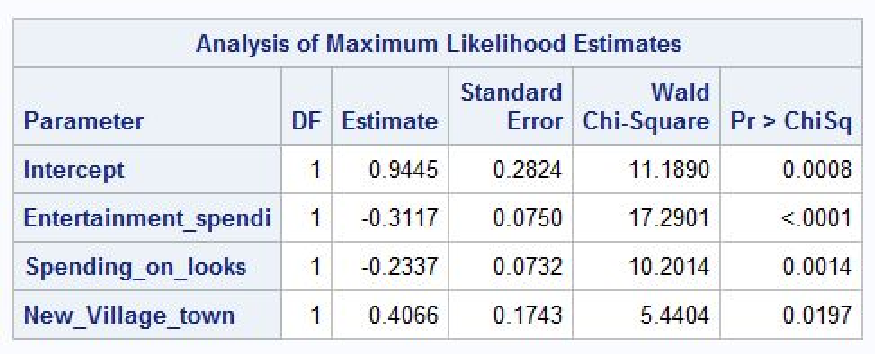
Fig 6.



The FREQ Procedure

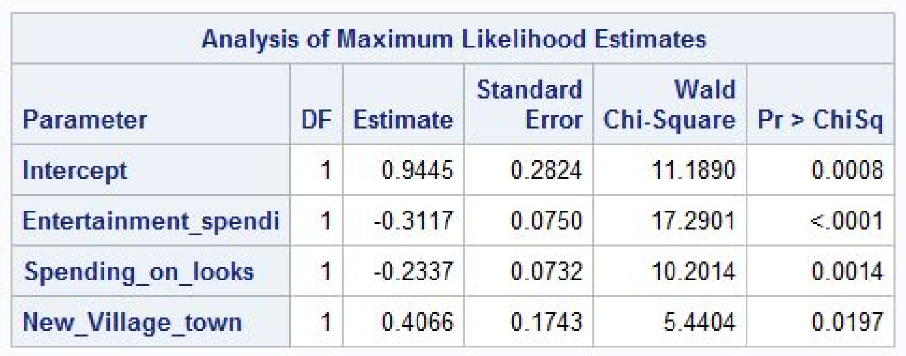
Selection Indicator				
Selected	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	247	25.00	247	25.00
1	741	75.00	988	100.00

Fig 7.



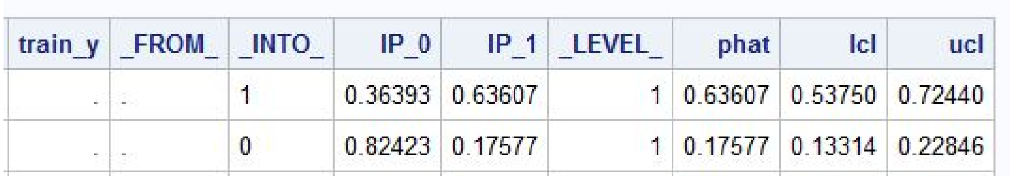
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.9445	0.2824	11.1890	0.0008
Entertainment_spendi	1	-0.3117	0.0750	17.2901	<.0001
Spending_on_looks	1	-0.2337	0.0732	10.2014	0.0014
New_Village_town	1	0.4066	0.1743	5.4404	0.0197

Fig 8.



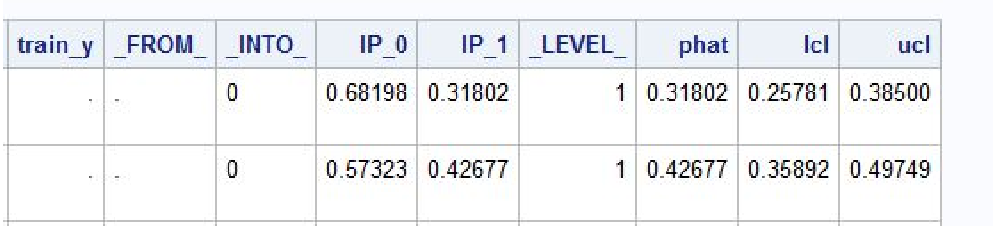
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.9445	0.2824	11.1890	0.0008
Entertainment_spendi	1	-0.3117	0.0750	17.2901	<.0001
Spending_on_looks	1	-0.2337	0.0732	10.2014	0.0014
New_Village_town	1	0.4066	0.1743	5.4404	0.0197

Fig 9.



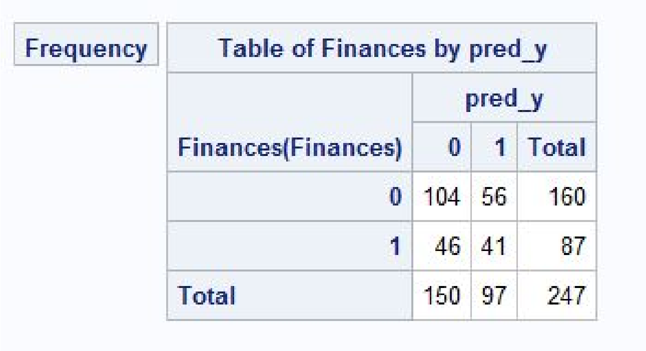
train_y	_FROM_	_INTO_	IP_0	IP_1	_LEVEL_	phat	lcl	ucl
.	.	1	0.36393	0.63607	1	0.63607	0.53750	0.72440
.	.	0	0.82423	0.17577	1	0.17577	0.13314	0.22846

Fig 10.



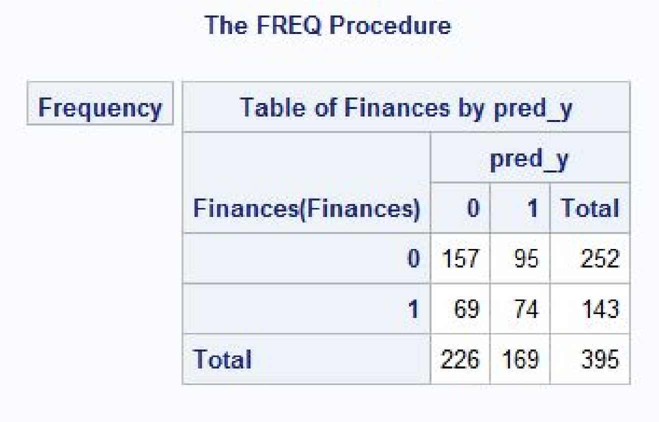
train_y	_FROM_	_INTO_	IP_0	IP_1	_LEVEL_	phat	lcl	ucl
.	.	0	0.68198	0.31802	1	0.31802	0.25781	0.38500
.	.	0	0.57323	0.42677	1	0.42677	0.35892	0.49749

Fig 11



Frequency	Table of Finances by pred_y			
		pred_y		
	Finances(Finances)	0	1	Total
	0	104	56	160
	1	46	41	87
	Total	150	97	247

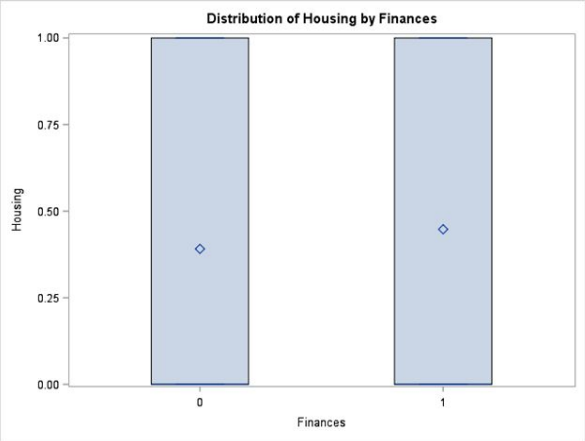
Fig 12.

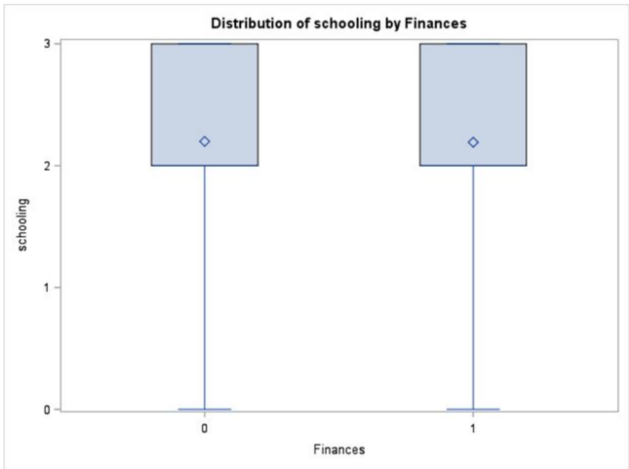


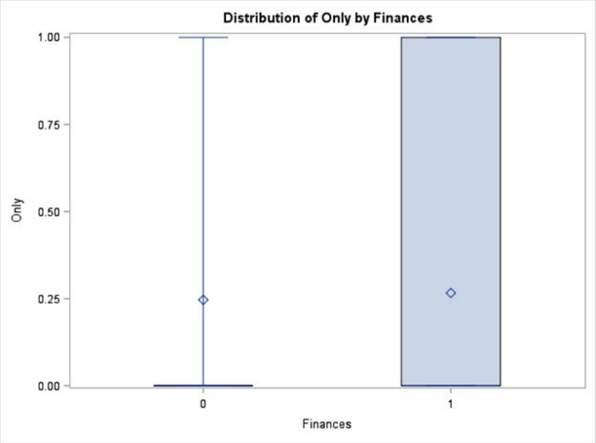
The FREQ Procedure

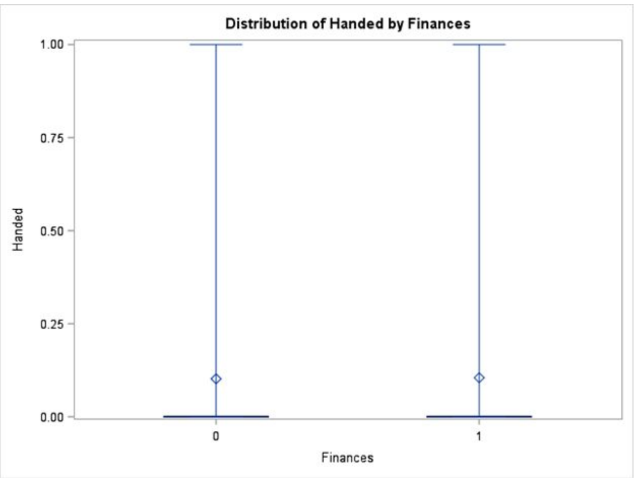
Frequency	Table of Finances by pred_y			
		pred_y		
	Finances(Finances)	0	1	Total
	0	157	95	252
	1	69	74	143
	Total	226	169	395

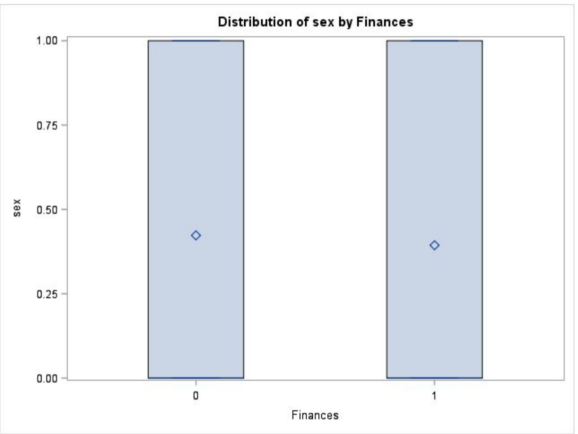
Rushabh Shah section

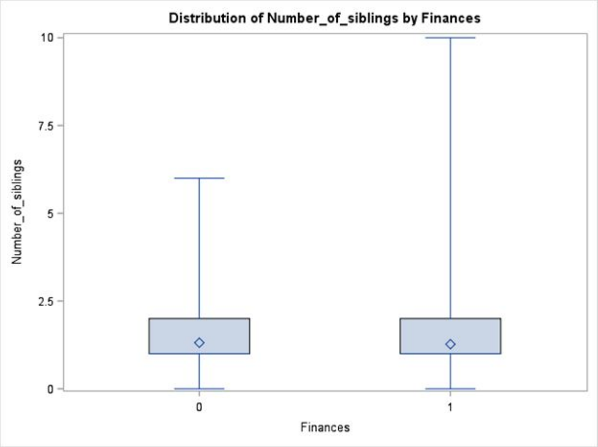


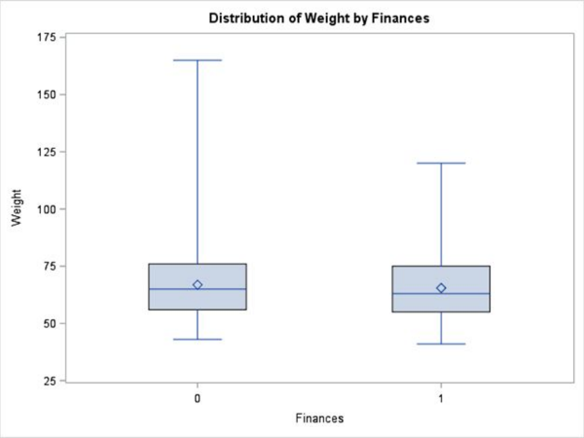


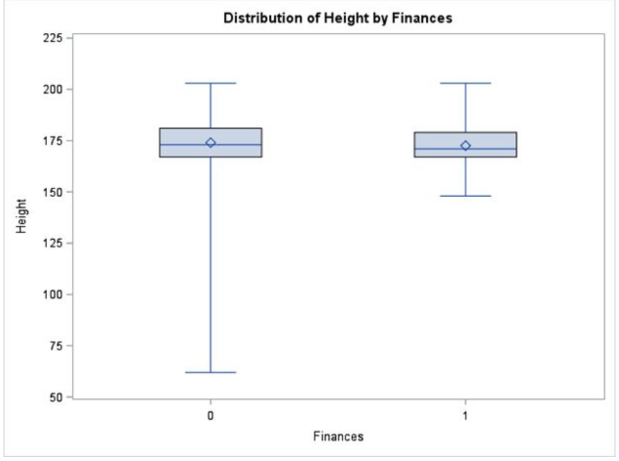

Distribution of Only by Finances
Only
1.00
0.75
0.50
0.25
0.00
0
1
Finances

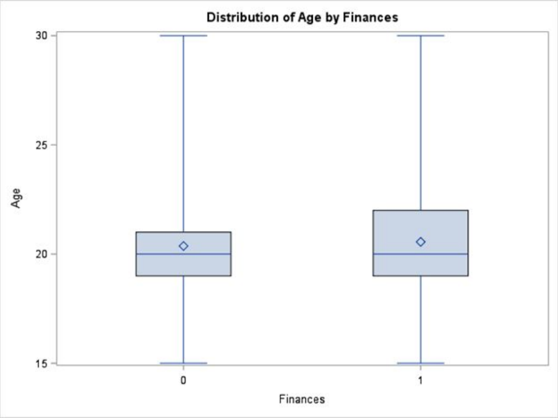

Distribution of Handed by Finances
Handed
1.00
0.75
0.50
0.25
0.00
0
1
Finances


Distribution of sex by Finances
sex
1.00
0.75
0.50
0.25
0.00
0
1
Finances

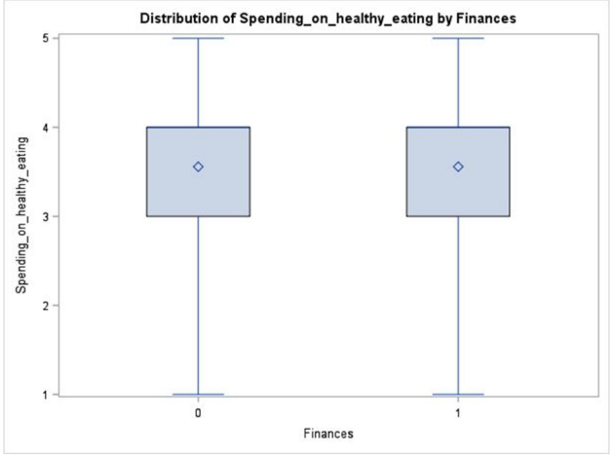

Distribution of Number_of_siblings by Finances
Number_of_siblings
10
7.5
5
2.5
0
0
1
Finances


Distribution of Weight by Finances
Weight
175
150
125
100
75
50
25
0
1
Finances

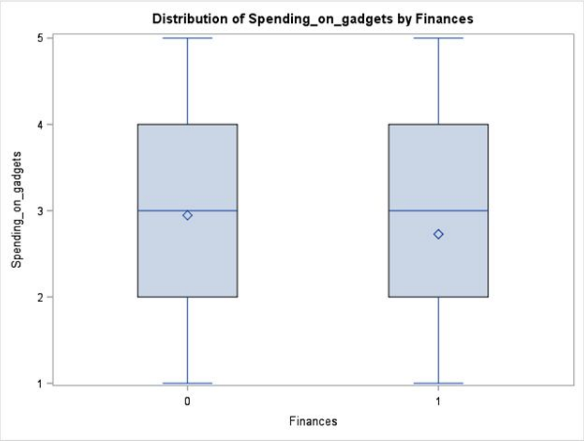

Distribution of Height by Finances
Height
225
200
175
150
125
100
75
50
0
1
Finances



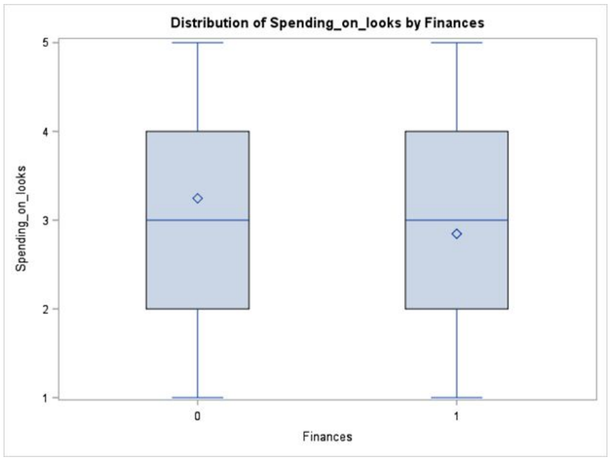
Distribution of Age by Finances



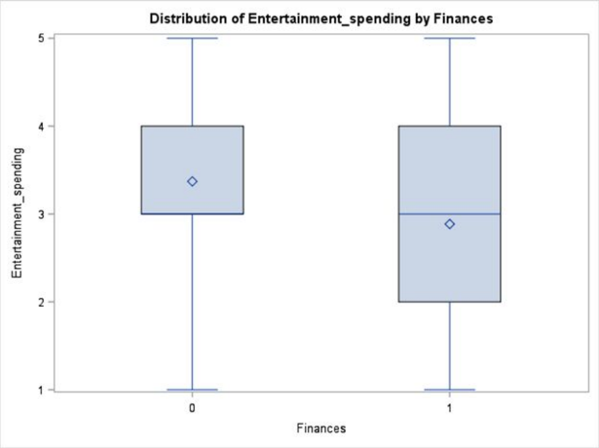
Distribution of Spending_on_healthy_eating by Finances



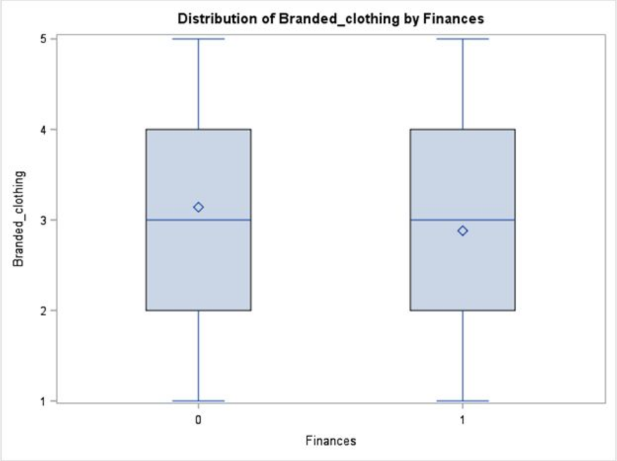
Distribution of Spending_on_gadgets by Finances



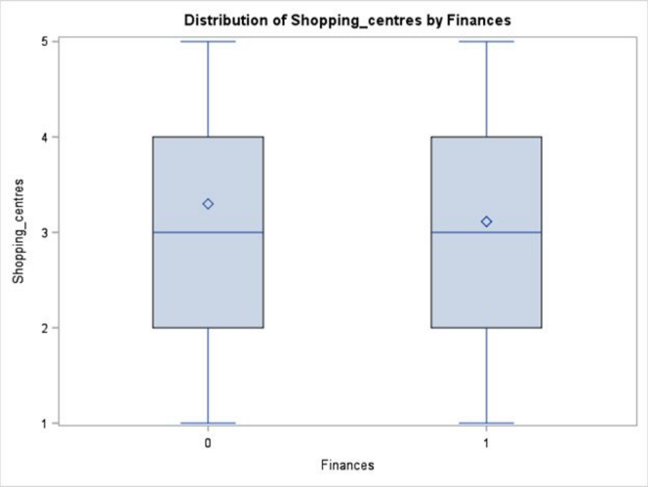
Distribution of Spending_on_looks by Finances



Distribution of Entertainment_spending by Finances



Distribution of Branded_clothing by Finances



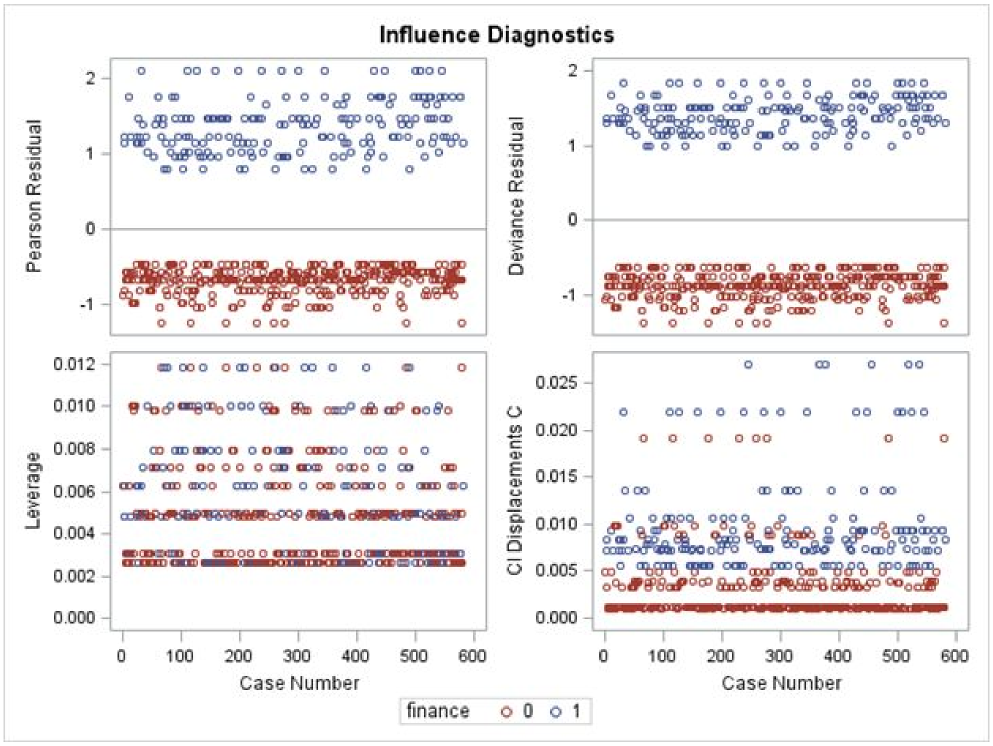
Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-2.7857	1.7758	2.4610	0.1167
Housing	1	0.0312	0.1787	0.0305	0.8614
Urban	1	0.3841	0.1903	4.0737	0.0436
schooling	1	0.0297	0.1347	0.0487	0.8254
Only	1	-0.0655	0.1832	0.1279	0.7206
Handed	1	-0.1350	0.2315	0.3403	0.5597
sex	1	-0.1232	0.2145	0.3300	0.5656
Number_of_siblings	1	0.0598	0.0795	0.5669	0.4515
Weight	1	0.00414	0.00815	0.2580	0.6115
Height	1	0.0140	0.0107	1.7015	0.1921
Age	1	-0.0302	0.0297	1.0319	0.3097

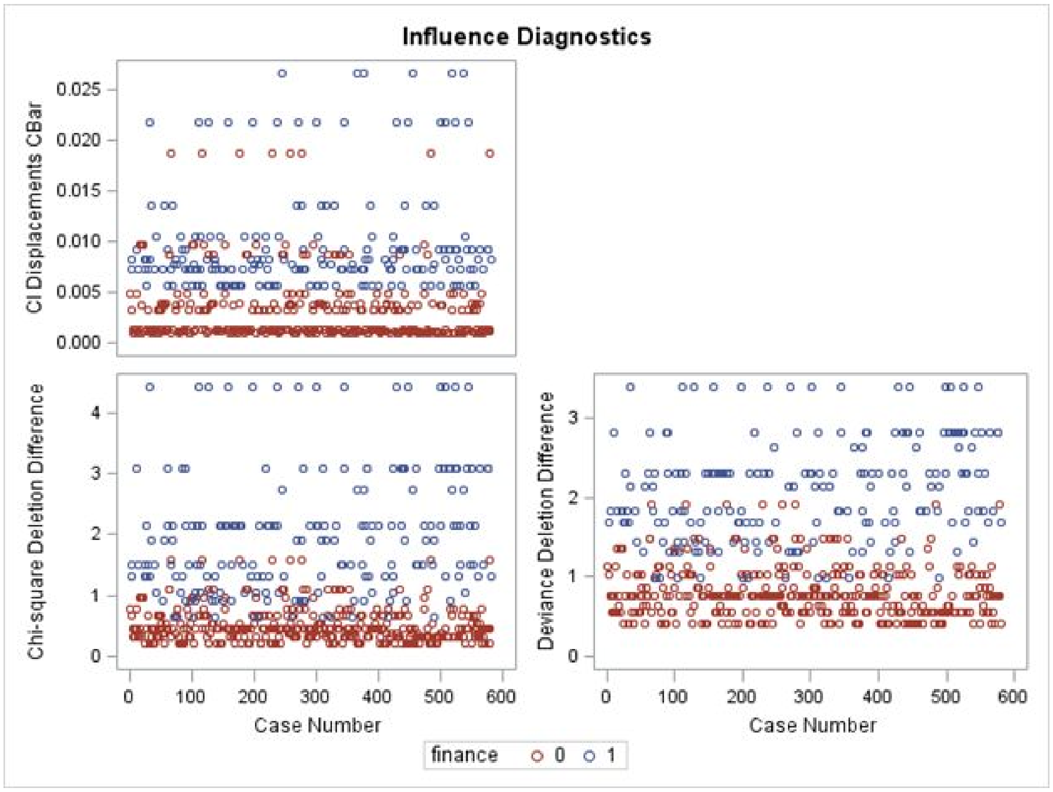
Spending_on_healthy_	1	-0.1274	0.0680	3.5125	0.0609
Spending_on_gadgets	1	0.00199	0.0641	0.0010	0.9753
Spending_on_looks	1	0.1910	0.0772	6.1198	0.0134
Entertainment_spendi	1	0.2698	0.0687	15.4397	<.0001
Branded_clothing	1	-0.0118	0.0648	0.0331	0.8557
Shopping_centres	1	0.0130	0.0636	0.0417	0.8383

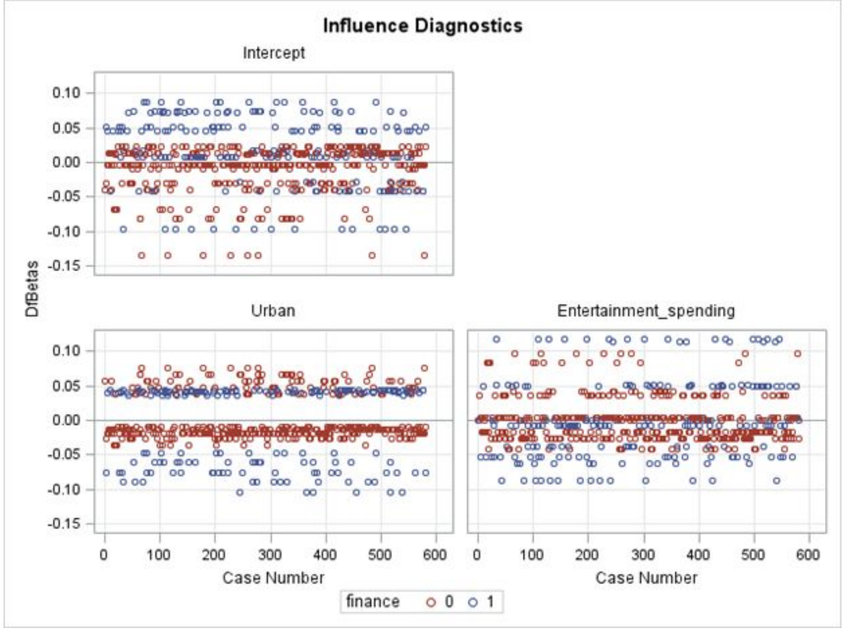
Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Housing	1.032	0.727	1.464
Urban	1.468	1.011	2.132
schooling	1.030	0.791	1.341
Only	0.937	0.654	1.341
Handed	0.874	0.555	1.375
sex	0.884	0.581	1.346
Number_of_siblings	1.062	0.909	1.241
Weight	1.004	0.988	1.020
Height	1.014	0.993	1.036
Age	0.970	0.915	1.028
Spending_on_healthy_	0.880	0.771	1.006

Spending_on_gadgets	1.002	0.884	1.136
Spending_on_looks	1.211	1.040	1.408
Entertainment_spendi	1.310	1.145	1.498
Branded_clothing	0.988	0.870	1.122
Shopping_centres	1.013	0.894	1.148

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	561.679	550.196
SC	565.734	562.359
-2 Log L	559.679	544.196

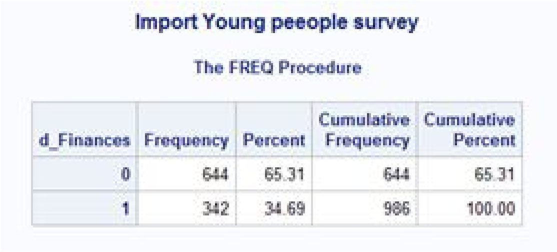

Influence Diagnostics
Pearson Residual
2
1
0
-1
Deviance Residual
2
1
0
-1
Leverage
0.012
0.010
0.008
0.006
0.004
0.002
0.000
CI Displacements C
0.025
0.020
0.015
0.010
0.005
0.000
0
100
200
300
400
500
600
Case Number
finance ○ 0 ○ 1


Influence Diagnostics
CI Displacements CBar
0.025
0.020
0.015
0.010
0.005
0.000
Chi-square Deletion Difference
4
3
2
1
0
Deviance Deletion Difference
3
2
1
0
0
100
200
300
400
500
600
Case Number
finance ○ 0 ○ 1


Influence Diagnostics
Intercept
Urban
Entertainment_spending
DfBetas
0.10
0.05
0.00
-0.05
-0.10
-0.15
0
100
200
300
400
500
600
Case Number
finance ○ 0 ○ 1

Valentine Silvester Correia section:

Frequency of savers/not savers.

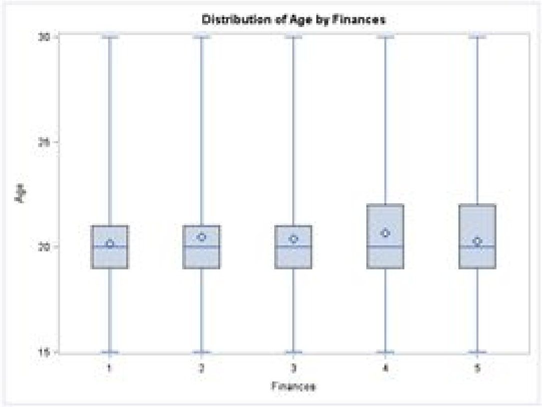


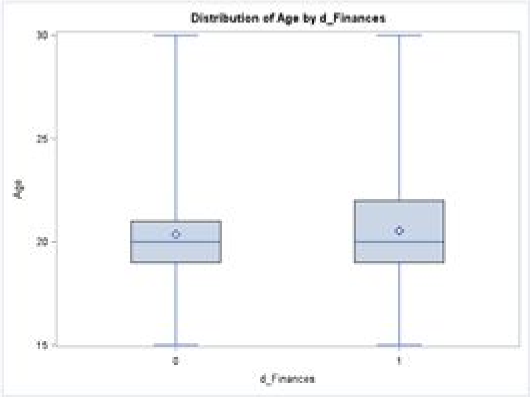
Import Young peeople survey

The FREQ Procedure

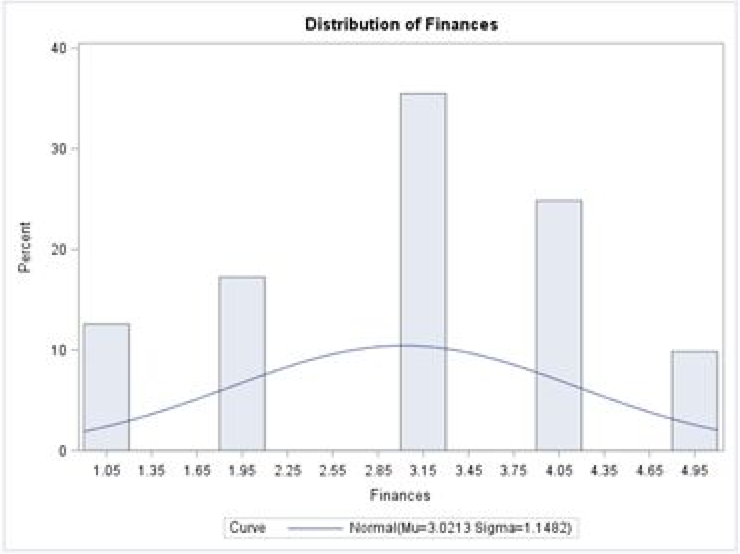
d_Finances	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	644	65.31	644	65.31
1	342	34.69	986	100.00

Boxplots for Finances before and after converting it to binary.

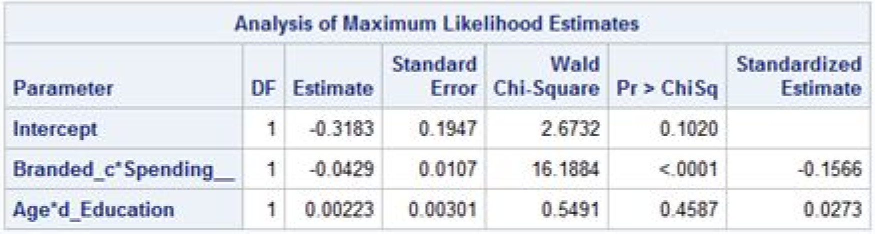




Histogram for variable Finances before converting it to binary.

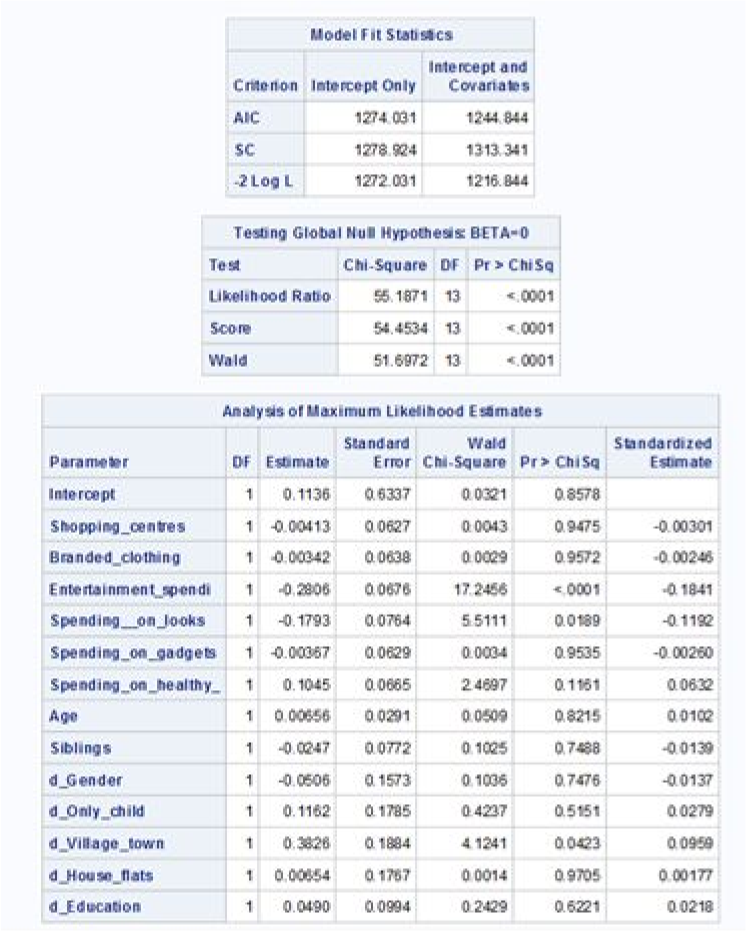


Checking for interactions.



Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-0.3183	0.1947	2.6732	0.1020	
Branded_c*Spending__	1	-0.0429	0.0107	16.1884	<.0001	-0.1566
Age*d_Education	1	0.00223	0.00301	0.5491	0.4587	0.0273

Verifying the full model for the whole dataset.

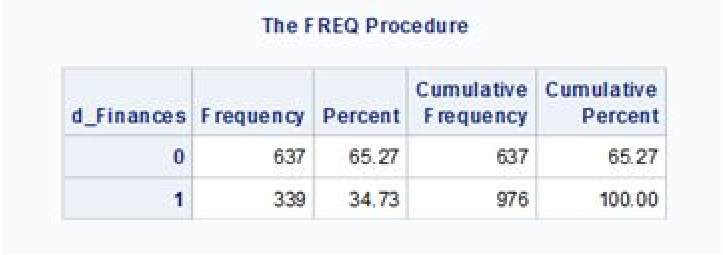


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1274.031	1244.844
SC	1278.924	1313.341
-2 Log L	1272.031	1216.844

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	55.1871	13	<.0001
Score	54.4534	13	<.0001
Wald	51.6972	13	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	0.1136	0.6337	0.0321	0.8578	
Shopping_centres	1	-0.00413	0.0627	0.0043	0.9475	-0.00301
Branded_clothing	1	-0.00342	0.0638	0.0029	0.9572	-0.00246
Entertainment_spendi	1	-0.2806	0.0676	17.2456	<.0001	-0.1841
Spending__on_looks	1	-0.1793	0.0764	5.5111	0.0189	-0.1192
Spending_on_gadgets	1	-0.00367	0.0629	0.0034	0.9535	-0.00260
Spending_on_healthy_	1	0.1045	0.0665	2.4697	0.1161	0.0632
Age	1	0.00656	0.0291	0.0509	0.8215	0.0102
Siblings	1	-0.0247	0.0772	0.1025	0.7488	-0.0139
d_Gender	1	-0.0506	0.1573	0.1036	0.7476	-0.0137
d_Only_child	1	0.1162	0.1785	0.4237	0.5151	0.0279
d_Village_town	1	0.3826	0.1884	4.1241	0.0423	0.0959
d_House_flats	1	0.00654	0.1767	0.0014	0.9705	0.00177
d_Education	1	0.0490	0.0994	0.2429	0.6221	0.0218

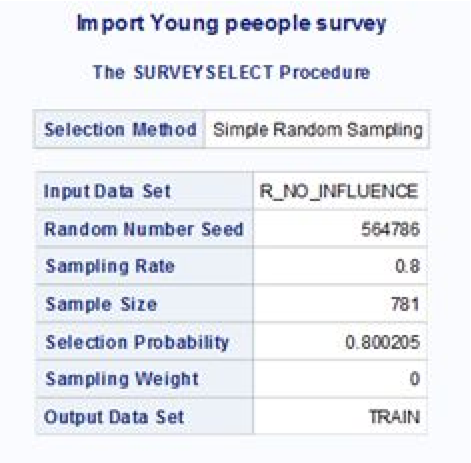
Frequency after checking Multicollinearity, Influential Point and Outliers.



The FREQ Procedure

d_Finances	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	637	65.27	637	65.27
1	339	34.73	976	100.00

Splitting the data into Training and Testing data.



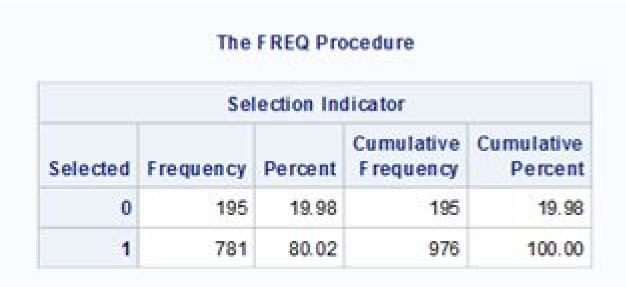
Import Young peeople survey

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	R_NO_INFLUENCE
Random Number Seed	564786
Sampling Rate	0.8
Sample Size	781
Selection Probability	0.800205
Sampling Weight	0
Output Data Set	TRAIN

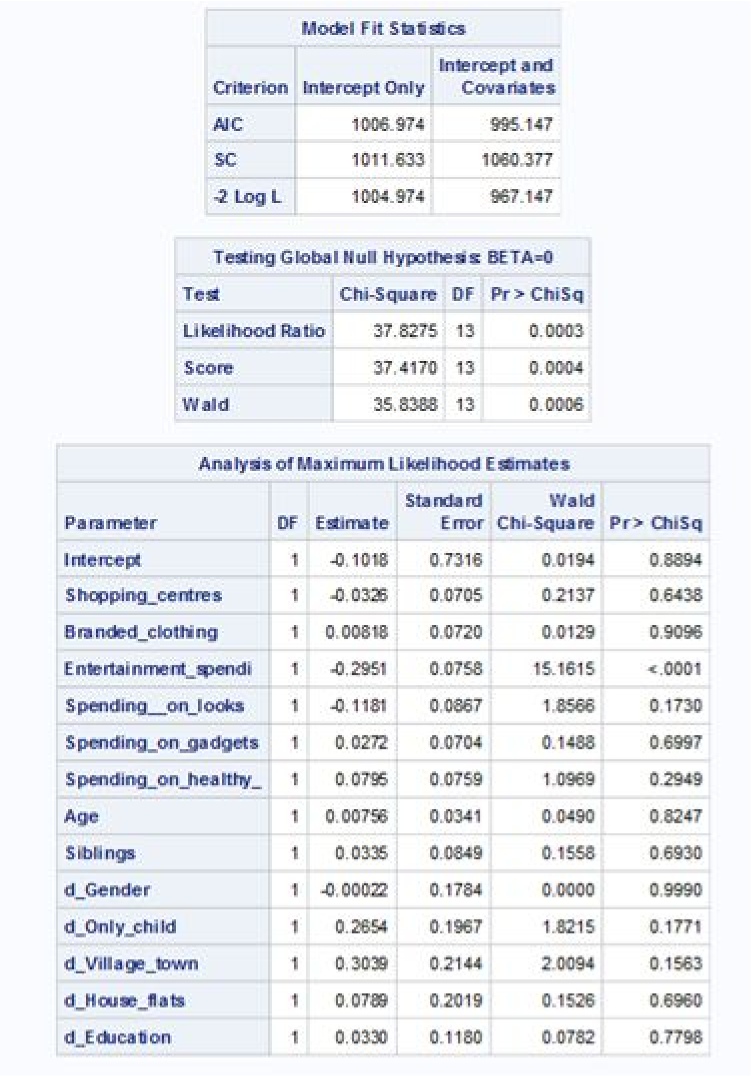
Verify if the data was split properly.



The FREQ Procedure

Selection Indicator				
Selected	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	195	19.98	195	19.98
1	781	80.02	976	100.00

Verifying the full model for the training dataset.

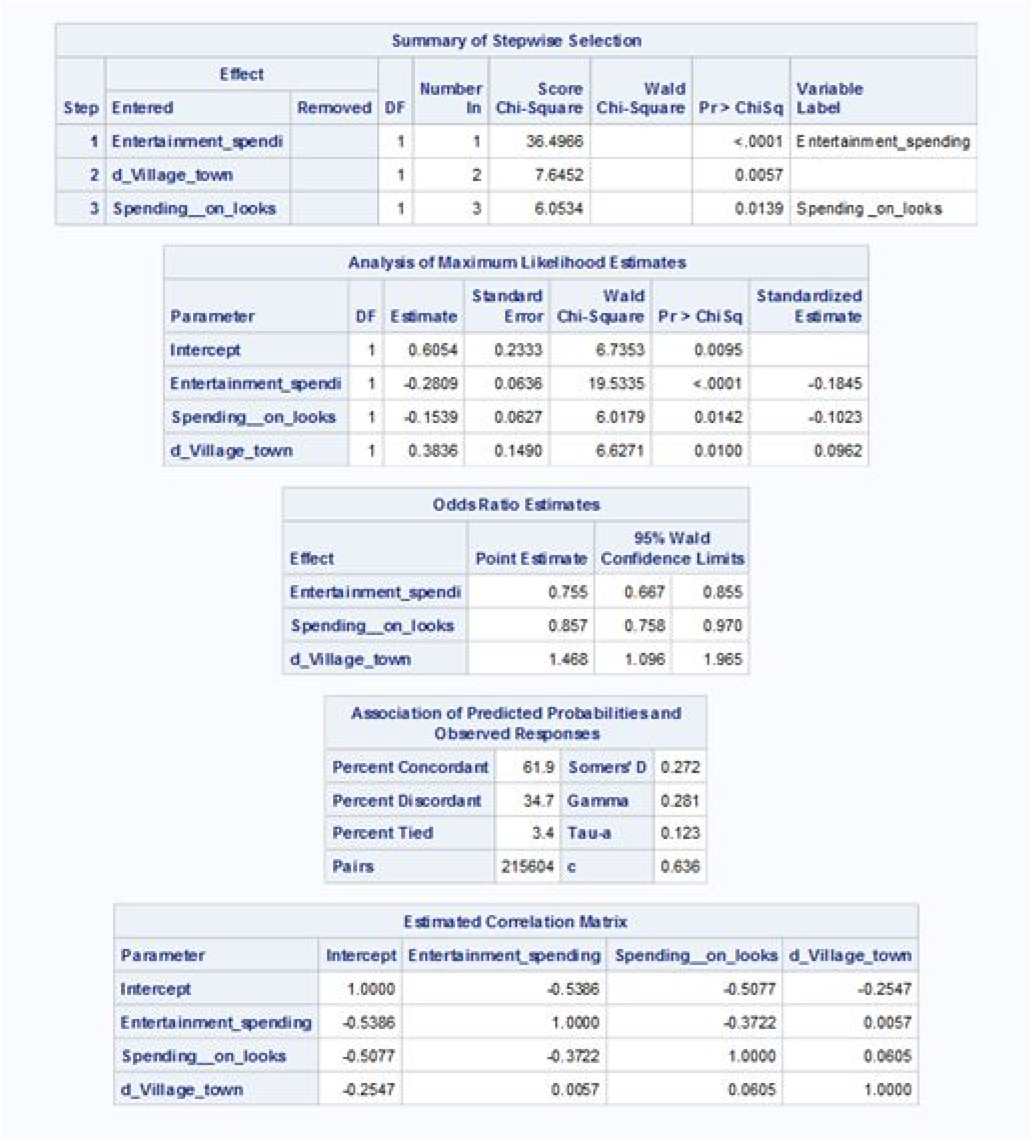


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1006.974	995.147
SC	1011.633	1060.377
-2 Log L	1004.974	967.147

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	37.8275	13	0.0003
Score	37.4170	13	0.0004
Wald	35.8388	13	0.0006

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.1018	0.7316	0.0194	0.8894
Shopping_centres	1	-0.0326	0.0705	0.2137	0.6438
Branded_clothing	1	0.00818	0.0720	0.0129	0.9096
Entertainment_spendi	1	-0.2951	0.0758	15.1615	<.0001
Spending__on_looks	1	-0.1181	0.0867	1.8566	0.1730
Spending_on_gadgets	1	0.0272	0.0704	0.1488	0.6997
Spending_on_healthy_	1	0.0795	0.0759	1.0969	0.2949
Age	1	0.00756	0.0341	0.0490	0.8247
Siblings	1	0.0335	0.0849	0.1558	0.6930
d_Gender	1	-0.00022	0.1784	0.0000	0.9990
d_Only_child	1	0.2654	0.1967	1.8215	0.1771
d_Village_town	1	0.3039	0.2144	2.0094	0.1563
d_House_flats	1	0.0789	0.2019	0.1526	0.6960
d_Education	1	0.0330	0.1180	0.0782	0.7798

Using Model selection method: Stepwise.



Summary of Stepwise Selection

	Effect							
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	Entertainment_spendi		1	1	36.4966		<.0001	Entertainment_spending
2	d_Village_town		1	2	7.6452		0.0057	
3	Spending__on_looks		1	3	6.0534		0.0139	Spending _on_looks

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	0.6054	0.2333	6.7353	0.0095	
Entertainment_spendi	1	-0.2809	0.0636	19.5335	<.0001	-0.1845
Spending__on_looks	1	-0.1539	0.0627	6.0179	0.0142	-0.1023
d_Village_town	1	0.3836	0.1490	6.6271	0.0100	0.0962

Odds Ratio Estimates

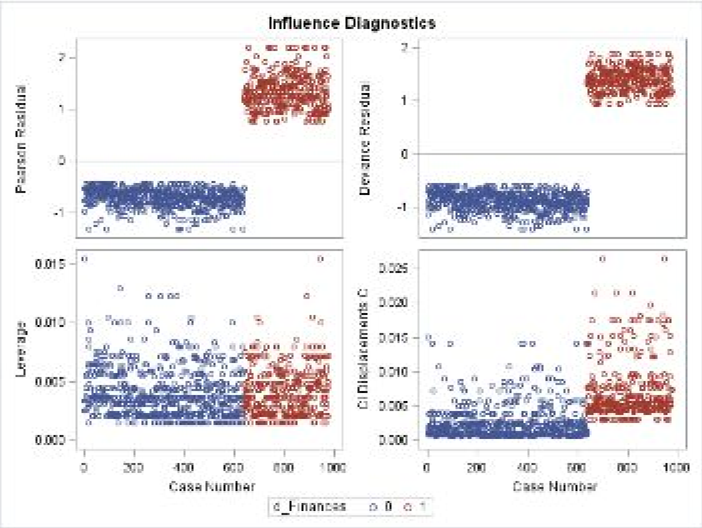
Effect	Point Estimate	95% Wald Confidence Limits	
Entertainment_spendi	0.755	0.667	0.855
Spending__on_looks	0.857	0.758	0.970
d_Village_town	1.468	1.096	1.965

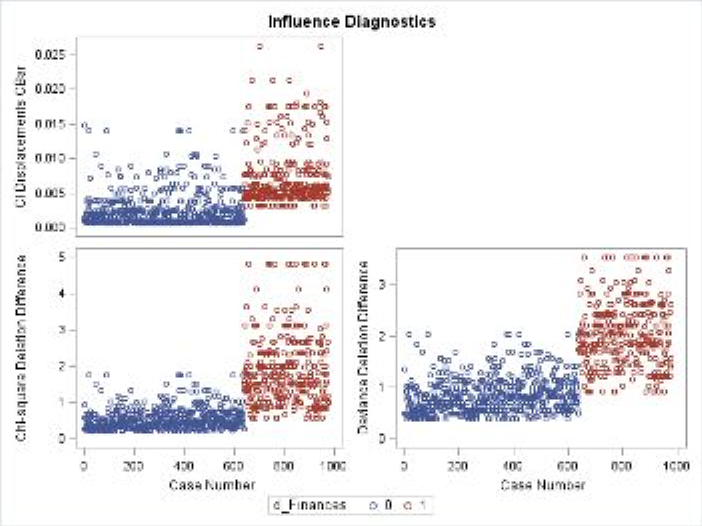
Association of Predicted Probabilities and Observed Responses

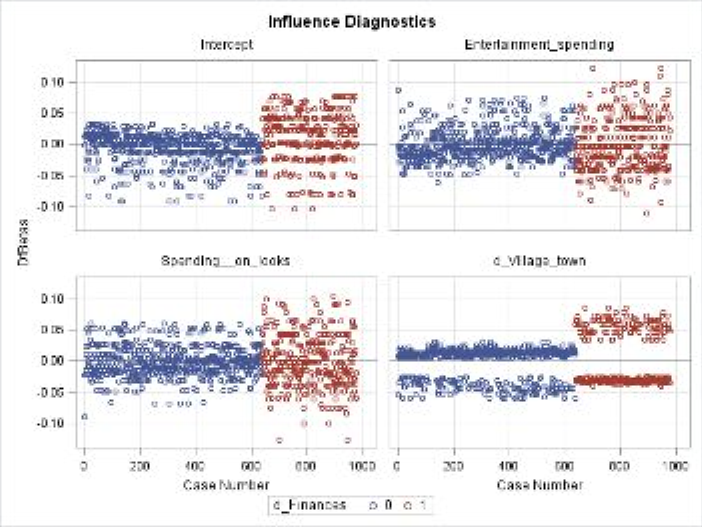
Percent Concordant	61.9	Somers' D	0.272
Percent Discordant	34.7	Gamma	0.281
Percent Tied	3.4	Tau-a	0.123
Pairs	215604	c	0.636

Estimated Correlation Matrix

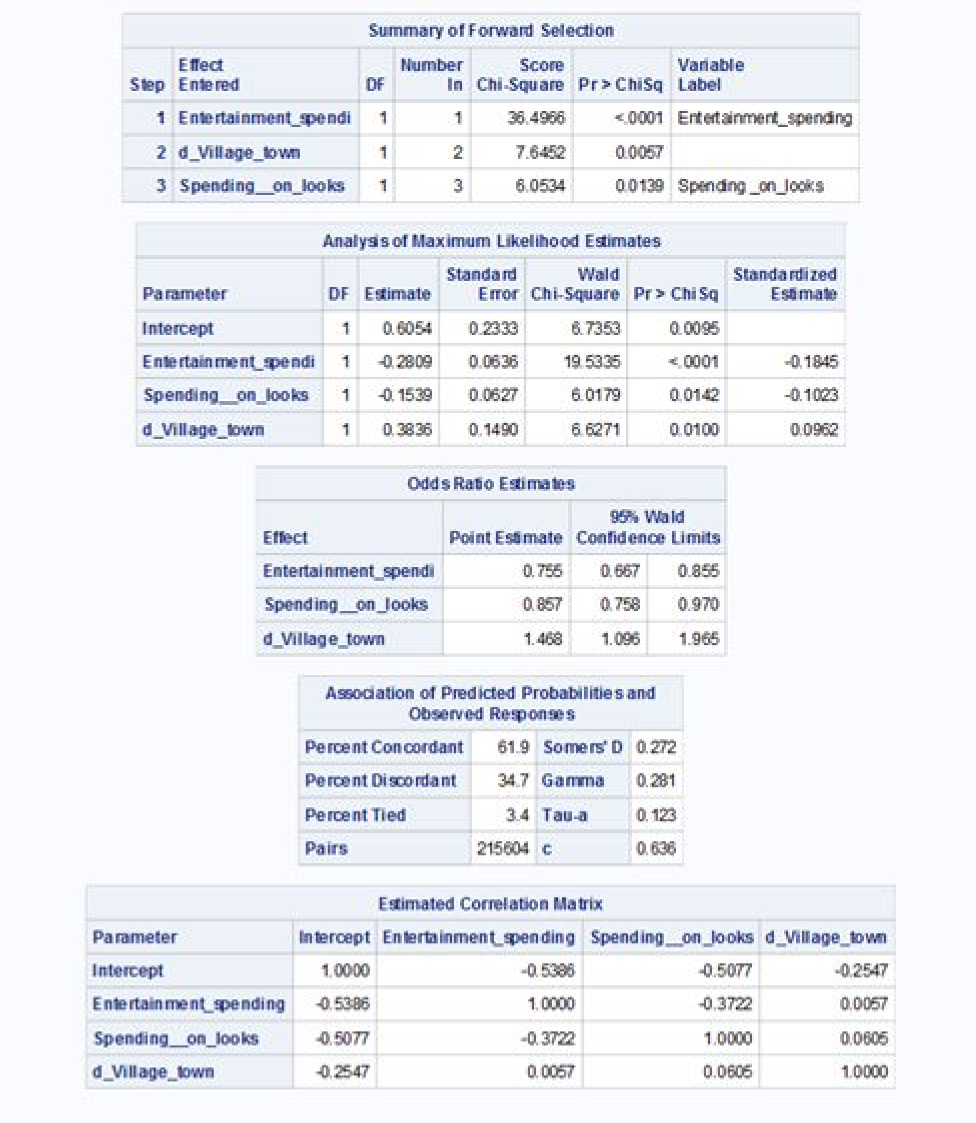
Parameter	Intercept	Entertainment_spending	Spending__on_looks	d_Village_town
Intercept	1.0000	-0.5386	-0.5077	-0.2547
Entertainment_spending	-0.5386	1.0000	-0.3722	0.0057
Spending__on_looks	-0.5077	-0.3722	1.0000	0.0605
d_Village_town	-0.2547	0.0057	0.0605	1.0000


Influence Diagnostics
Pearson Residual
Deviance Residual
Leverage
CI Displacements C
Case Number
Case Number
0
200
400
600
800
1000
2
1
0
-1
0.015
0.010
0.005
0.000
0.025
0.020
0.015
0.010
0.005
0.000
d_Finances ○ 0 ○ 1


Influence Diagnostics
CI Displacements CBar
Chi-square Deletion Difference
Deviance Deletion Difference
Case Number
Case Number
0.025
0.020
0.015
0.010
0.005
0.000
5
4
3
2
1
0
0
200
400
600
800
1000
d_Finances ○ 0 ○ 1


Influence Diagnostics
Intercept
Entertainment_spending
Spending__on_looks
d_Village_town
DfBetas
0.10
0.05
0.00
-0.05
-0.10
Case Number
Case Number
0
200
400
600
800
1000
d_Finances ○ 0 ○ 1

Using Model selection method: Backward.



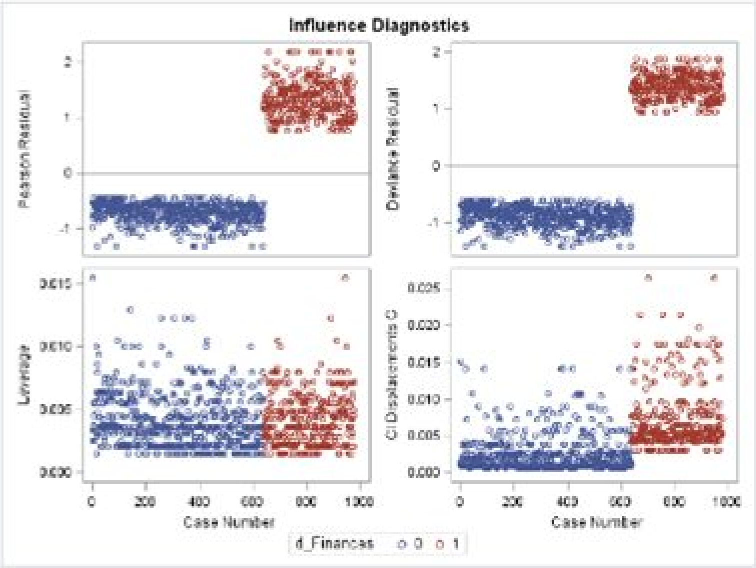
Summary of Forward Selection						
Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Variable Label
1	Entertainment_spendi	1	1	36.4966	<.0001	Entertainment_spending
2	d_Village_town	1	2	7.6452	0.0057	
3	Spending__on_looks	1	3	6.0534	0.0139	Spending _on_looks

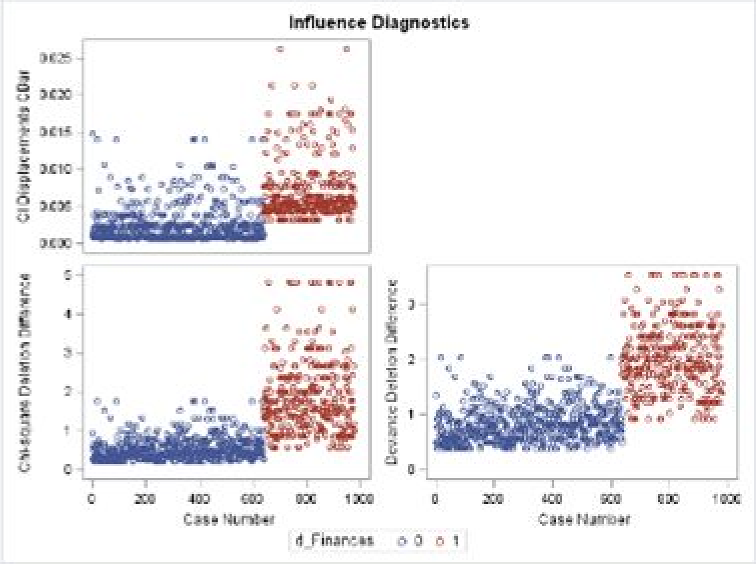
Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	0.6054	0.2333	6.7353	0.0095	
Entertainment_spendi	1	-0.2809	0.0636	19.5335	<.0001	-0.1845
Spending__on_looks	1	-0.1539	0.0627	6.0179	0.0142	-0.1023
d_Village_town	1	0.3836	0.1490	6.6271	0.0100	0.0962

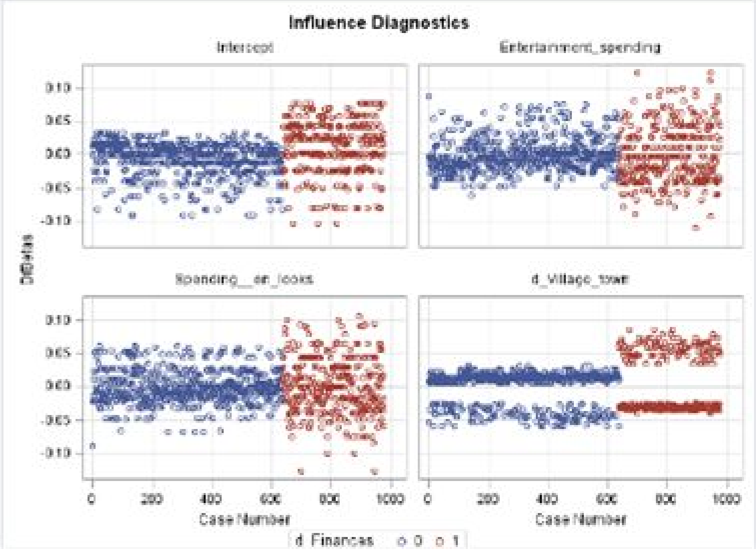
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Entertainment_spendi	0.755	0.667	0.855
Spending__on_looks	0.857	0.758	0.970
d_Village_town	1.468	1.096	1.965

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	61.9	Somers' D	0.272
Percent Discordant	34.7	Gamma	0.281
Percent Tied	3.4	Tau-a	0.123
Pairs	215604	c	0.636

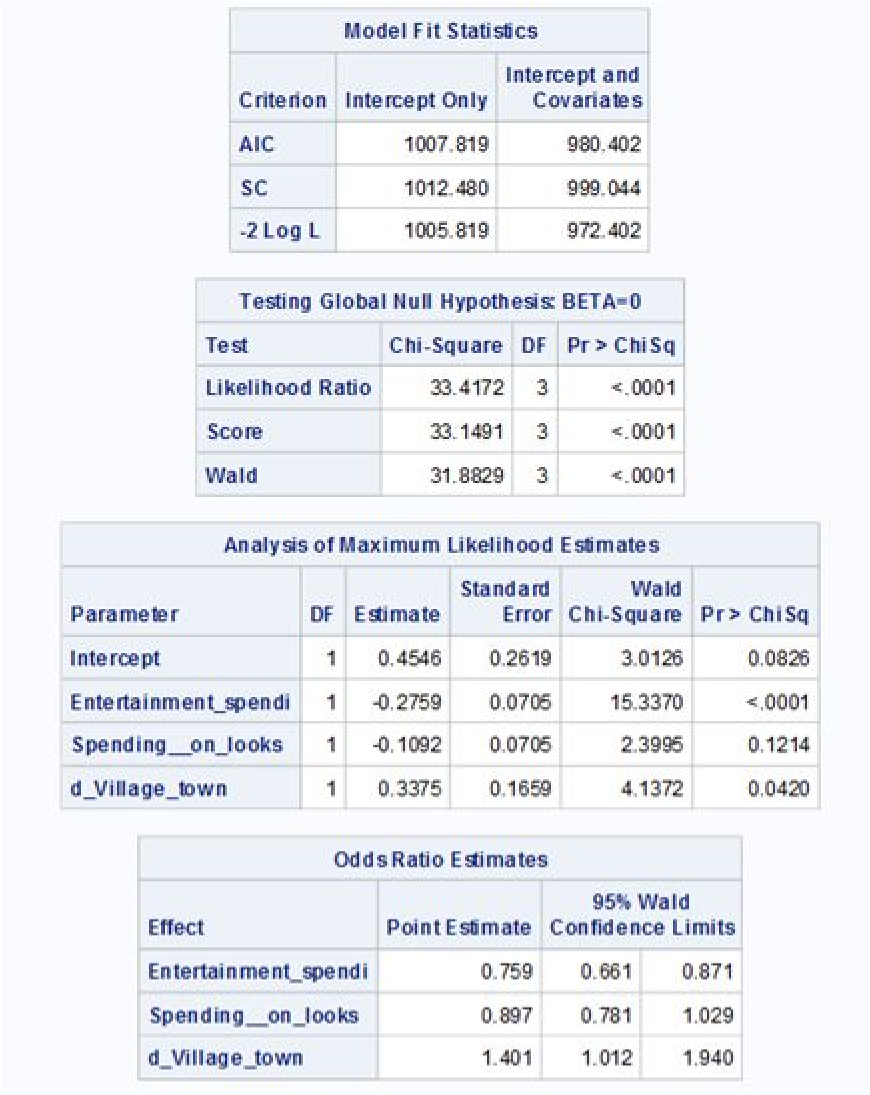
Estimated Correlation Matrix				
Parameter	Intercept	Entertainment_spending	Spending__on_looks	d_Village_town
Intercept	1.0000	-0.5386	-0.5077	-0.2547
Entertainment_spending	-0.5386	1.0000	-0.3722	0.0057
Spending__on_looks	-0.5077	-0.3722	1.0000	0.0605
d_Village_town	-0.2547	0.0057	0.0605	1.0000


Influence Diagnostics
Pearson Residual
Deviance Residual
Leverage
CI Displacements C
Case Number
Case Number
d_Financas ○ 0 ○ 1


Influence Diagnostics
CI Displacements CBar
Chi-square Deletion Difference
Deviance Deletion Difference
Case Number
Case Number
d_Financas ○ 0 ○ 1


Influence Diagnostics
Intercept
Entertainment_spending
Spending__on_looks
d_Village_town
DfBetas
Case Number
Case Number
d_Financas ○ 0 ○ 1

Find the final model, compute predicted value on training dataset and obtain the cut-off value.



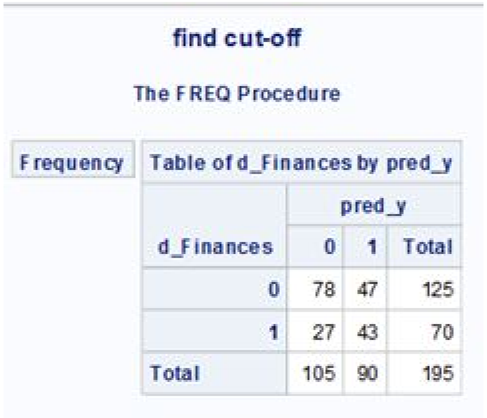
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1007.819	980.402
SC	1012.480	999.044
-2 Log L	1005.819	972.402

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	33.4172	3	<.0001
Score	33.1491	3	<.0001
Wald	31.8829	3	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.4546	0.2619	3.0126	0.0826
Entertainment_spendi	1	-0.2759	0.0705	15.3370	<.0001
Spending__on_looks	1	-0.1092	0.0705	2.3995	0.1214
d_Village_town	1	0.3375	0.1659	4.1372	0.0420

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Entertainment_spendi	0.759	0.661	0.871
Spending__on_looks	0.897	0.781	1.029
d_Village_town	1.401	1.012	1.940

Find cut-off value.



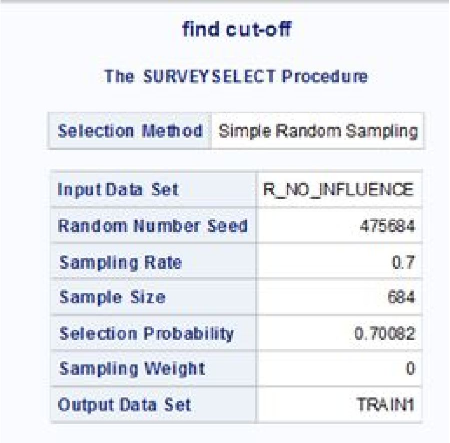
find cut-off

The FREQ Procedure

Frequency	Table of d_Finances by pred_y		
	pred_y		
d_Finances	0	1	Total
0	78	47	125
1	27	43	70
Total	105	90	195

Second Model: -

Training testing Data distribution.



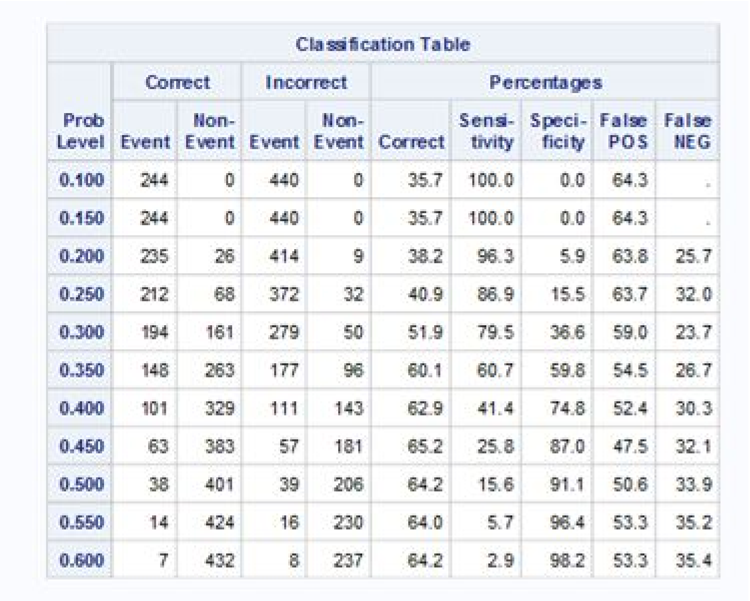
find cut-off

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

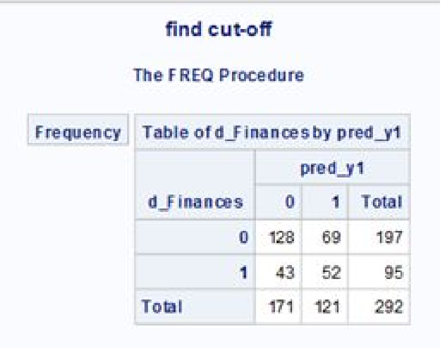
Input Data Set	R_NO_INFLUENCE
Random Number Seed	475684
Sampling Rate	0.7
Sample Size	684
Selection Probability	0.70082
Sampling Weight	0
Output Data Set	TRAIN1

Classification table for cut-off value.



Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.100	244	0	440	0	35.7	100.0	0.0	64.3	.
0.150	244	0	440	0	35.7	100.0	0.0	64.3	.
0.200	235	26	414	9	38.2	96.3	5.9	63.8	25.7
0.250	212	68	372	32	40.9	86.9	15.5	63.7	32.0
0.300	194	161	279	50	51.9	79.5	36.6	59.0	23.7
0.350	148	263	177	96	60.1	60.7	59.8	54.5	26.7
0.400	101	329	111	143	62.9	41.4	74.8	52.4	30.3
0.450	63	383	57	181	65.2	25.8	87.0	47.5	32.1
0.500	38	401	39	206	64.2	15.6	91.1	50.6	33.9
0.550	14	424	16	230	64.0	5.7	96.4	53.3	35.2
0.600	7	432	8	237	64.2	2.9	98.2	53.3	35.4

Find cut-off value.



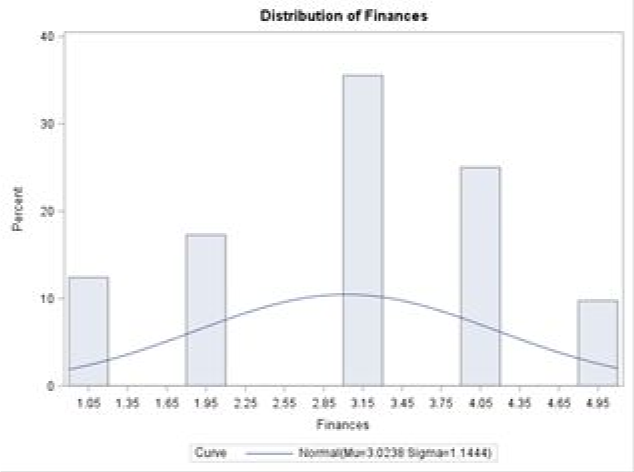
find cut-off

The FREQ Procedure

Frequency	Table of d_Finances by pred_y1			
		pred_y1		
	d_Finances	0	1	Total
	0	128	69	197
	1	43	52	95
	Total	171	121	292

Yesheng Qin section:

- **Response variable ‘Finances’ distribution**

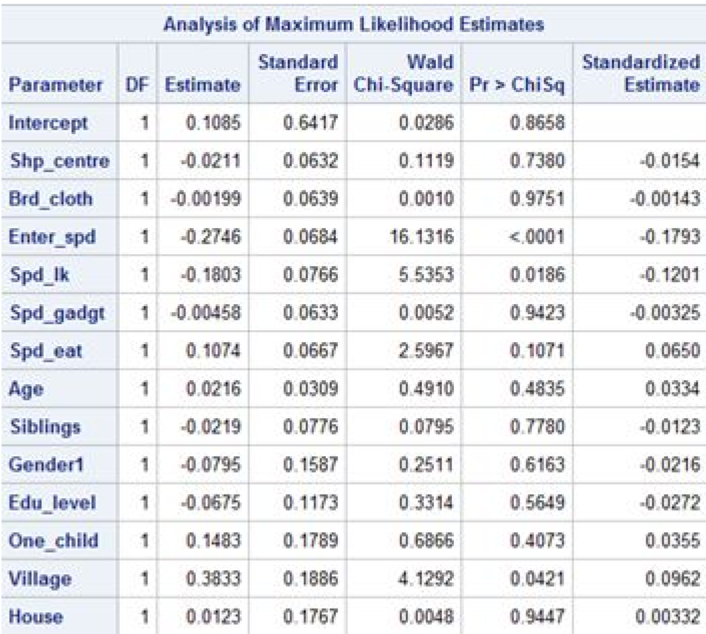


- **Dummy variable table**

Original	Dummy	Description
Gender	Gender1 (1,0)	1=male, 0=female
Education	Edu_level(1,2,3,4)	1=(Currently a Primary school pupil) or (Primary school) 2=(Secondary school) 3=(College/Bachelor degree) 4=(masters degree)
Only_child	One_child(1,0)	1=yes, 0=no
VorC	Village(1,0)	1=village, 0=city
HorB	House(1,0)	1=house, 0=block of flats

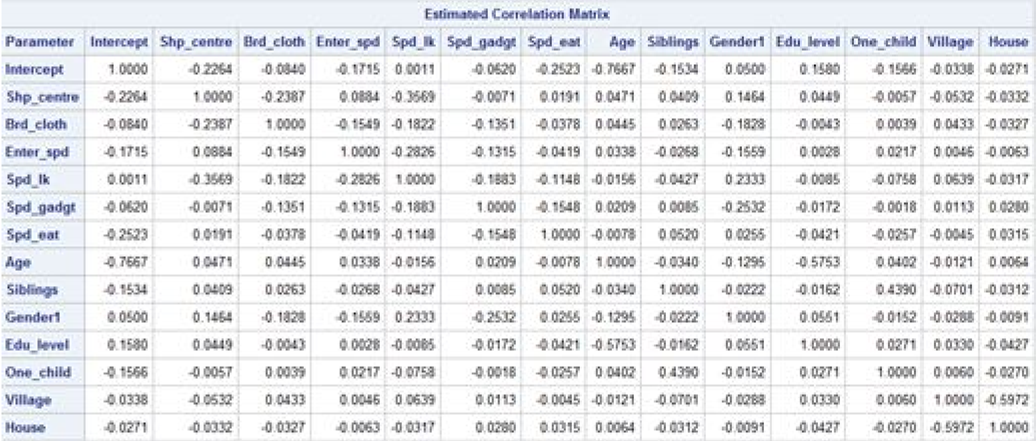
·

Full likelihood estimate table



Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	0.1085	0.6417	0.0286	0.8658	
Shp_centre	1	-0.0211	0.0632	0.1119	0.7380	-0.0154
Brd_cloth	1	-0.00199	0.0639	0.0010	0.9751	-0.00143
Enter_spd	1	-0.2746	0.0684	16.1316	<.0001	-0.1793
Spd_lk	1	-0.1803	0.0766	5.5353	0.0186	-0.1201
Spd_gadgt	1	-0.00458	0.0633	0.0052	0.9423	-0.00325
Spd_eat	1	0.1074	0.0667	2.5967	0.1071	0.0650
Age	1	0.0216	0.0309	0.4910	0.4835	0.0334
Siblings	1	-0.0219	0.0776	0.0795	0.7780	-0.0123
Gender1	1	-0.0795	0.1587	0.2511	0.6163	-0.0216
Edu_level	1	-0.0675	0.1173	0.3314	0.5649	-0.0272
One_child	1	0.1483	0.1789	0.6866	0.4073	0.0355
Village	1	0.3833	0.1886	4.1292	0.0421	0.0962
House	1	0.0123	0.1767	0.0048	0.9447	0.00332

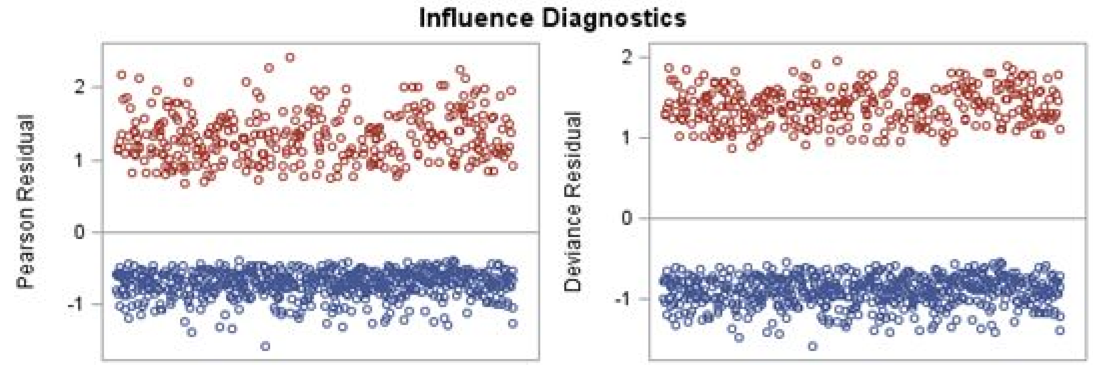
Full correlation table



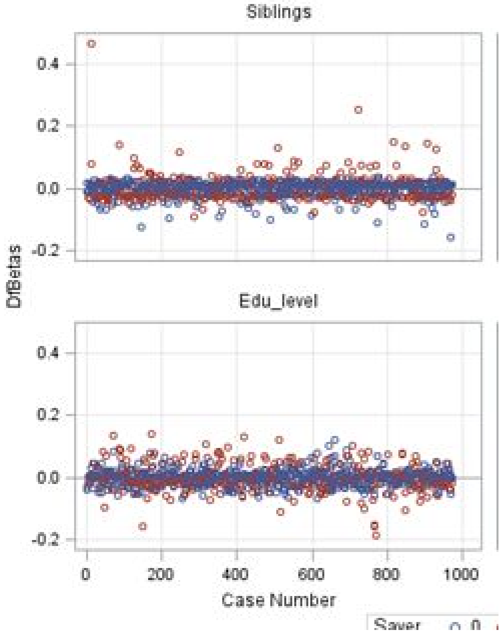
Estimated Correlation Matrix

Parameter	Intercept	Shp_centre	Brd_cloth	Enter_spd	Spd_lk	Spd_gadgt	Spd_eat	Age	Siblings	Gender1	Edu_level	One_child	Village	House
Intercept	1.0000	-0.2264	-0.0840	-0.1715	0.0011	-0.0620	-0.2523	-0.7667	-0.1534	0.0500	0.1580	-0.1566	-0.0338	-0.0271
Shp_centre	-0.2264	1.0000	-0.2387	0.0884	-0.3569	-0.0071	0.0191	0.0471	0.0409	0.1464	0.0449	-0.0057	-0.0532	-0.0332
Brd_cloth	-0.0840	-0.2387	1.0000	-0.1549	-0.1822	-0.1351	-0.0378	0.0445	0.0263	-0.1828	-0.0043	0.0039	0.0433	-0.0327
Enter_spd	-0.1715	0.0884	-0.1549	1.0000	-0.2826	-0.1315	-0.0419	0.0338	-0.0268	-0.1559	0.0028	0.0217	0.0046	-0.0063
Spd_lk	0.0011	-0.3569	-0.1822	-0.2826	1.0000	-0.1883	-0.1148	-0.0156	-0.0427	0.2333	-0.0085	-0.0758	0.0639	-0.0317
Spd_gadgt	-0.0620	-0.0071	-0.1351	-0.1315	-0.1883	1.0000	-0.1548	0.0209	0.0085	-0.2532	-0.0172	-0.0018	0.0113	0.0280
Spd_eat	-0.2523	0.0191	-0.0378	-0.0419	-0.1148	-0.1548	1.0000	-0.0078	0.0520	0.0255	-0.0421	-0.0257	-0.0045	0.0315
Age	-0.7667	0.0471	0.0445	0.0338	-0.0156	0.0209	-0.0078	1.0000	-0.0340	-0.1295	-0.5753	0.0402	-0.0121	0.0064
Siblings	-0.1534	0.0409	0.0263	-0.0268	-0.0427	0.0085	0.0520	-0.0340	1.0000	-0.0222	-0.0162	0.4390	-0.0701	-0.0312
Gender1	0.0500	0.1464	-0.1828	-0.1559	0.2333	-0.2532	0.0255	-0.1295	-0.0222	1.0000	0.0551	-0.0152	-0.0288	-0.0091
Edu_level	0.1580	0.0449	-0.0043	0.0028	-0.0085	-0.0172	-0.0421	-0.5753	-0.0162	0.0551	1.0000	0.0271	0.0330	-0.0427
One_child	-0.1566	-0.0057	0.0039	0.0217	-0.0758	-0.0018	-0.0257	0.0402	0.4390	-0.0152	0.0271	1.0000	0.0060	-0.0270
Village	-0.0338	-0.0532	0.0433	0.0046	0.0639	0.0113	-0.0045	-0.0121	-0.0701	-0.0288	0.0330	0.0060	1.0000	-0.5972
House	-0.0271	-0.0332	-0.0327	-0.0063	-0.0317	0.0280	0.0315	0.0064	-0.0312	-0.0091	-0.0427	-0.0270	-0.5972	1.0000

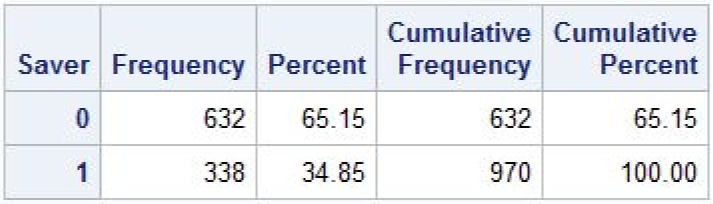
Full outliers



Full influential points

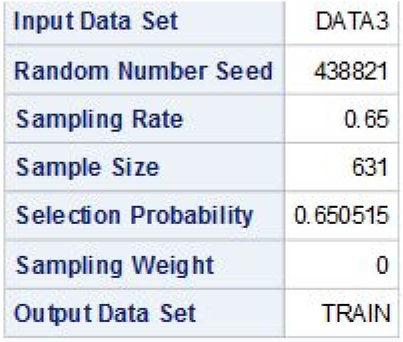


· **Full frequency table on Saver**



Saver	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	632	65.15	632	65.15
1	338	34.85	970	100.00

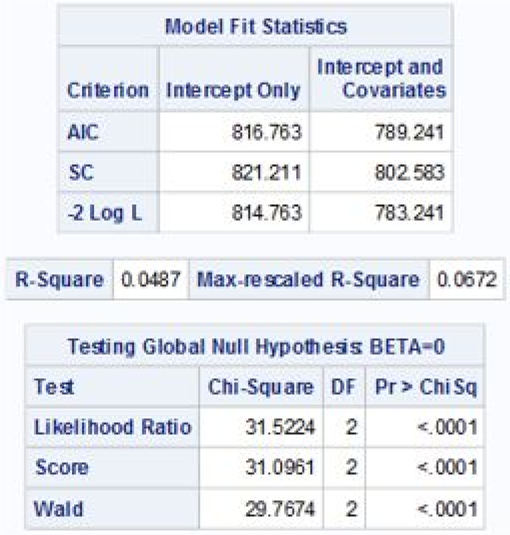
· **Training set table**



Input Data Set	DATA3
Random Number Seed	438821
Sampling Rate	0.65
Sample Size	631
Selection Probability	0.650515
Sampling Weight	0
Output Data Set	TRAIN

·

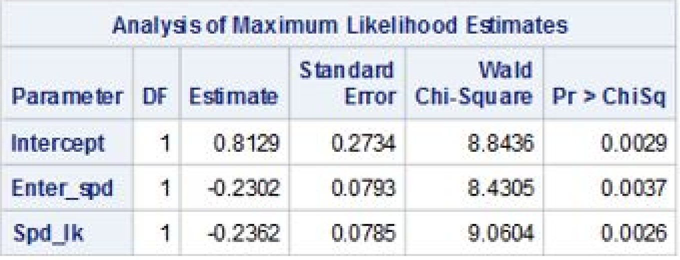
Stepwise



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	816.763	789.241
SC	821.211	802.583
-2 Log L	814.763	783.241

R-Square	0.0487	Max-rescaled R-Square	0.0672

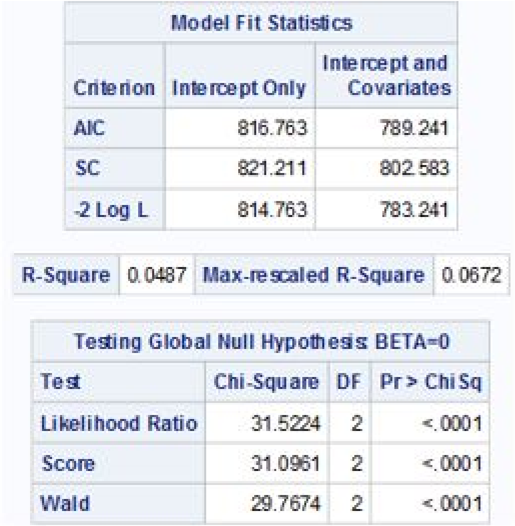
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	31.5224	2	<.0001
Score	31.0961	2	<.0001
Wald	29.7674	2	<.0001



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8129	0.2734	8.8436	0.0029
Enter_spd	1	-0.2302	0.0793	8.4305	0.0037
Spd_lk	1	-0.2362	0.0785	9.0604	0.0026

.

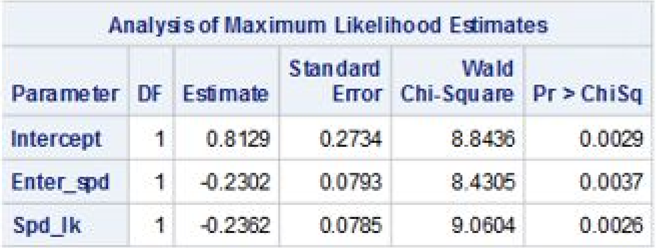
Backward



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	816.763	789.241
SC	821.211	802.583
-2 Log L	814.763	783.241

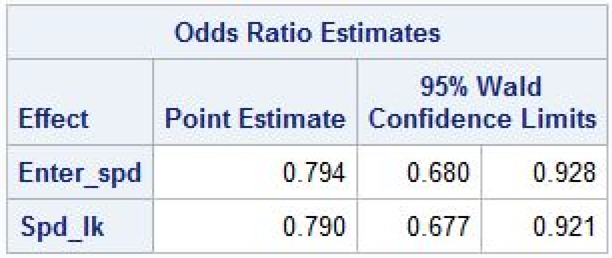
R-Square	0.0487	Max-rescaled R-Square	0.0672

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	31.5224	2	<.0001
Score	31.0961	2	<.0001
Wald	29.7674	2	<.0001



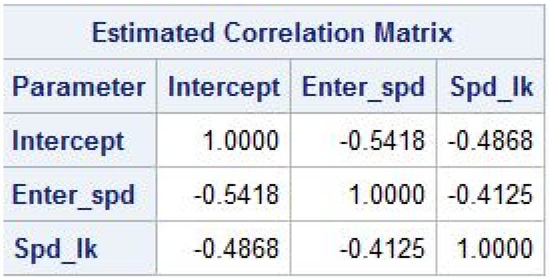
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8129	0.2734	8.8436	0.0029
Enter_spd	1	-0.2302	0.0793	8.4305	0.0037
Spd_lk	1	-0.2362	0.0785	9.0604	0.0026

- **Odd ratio**



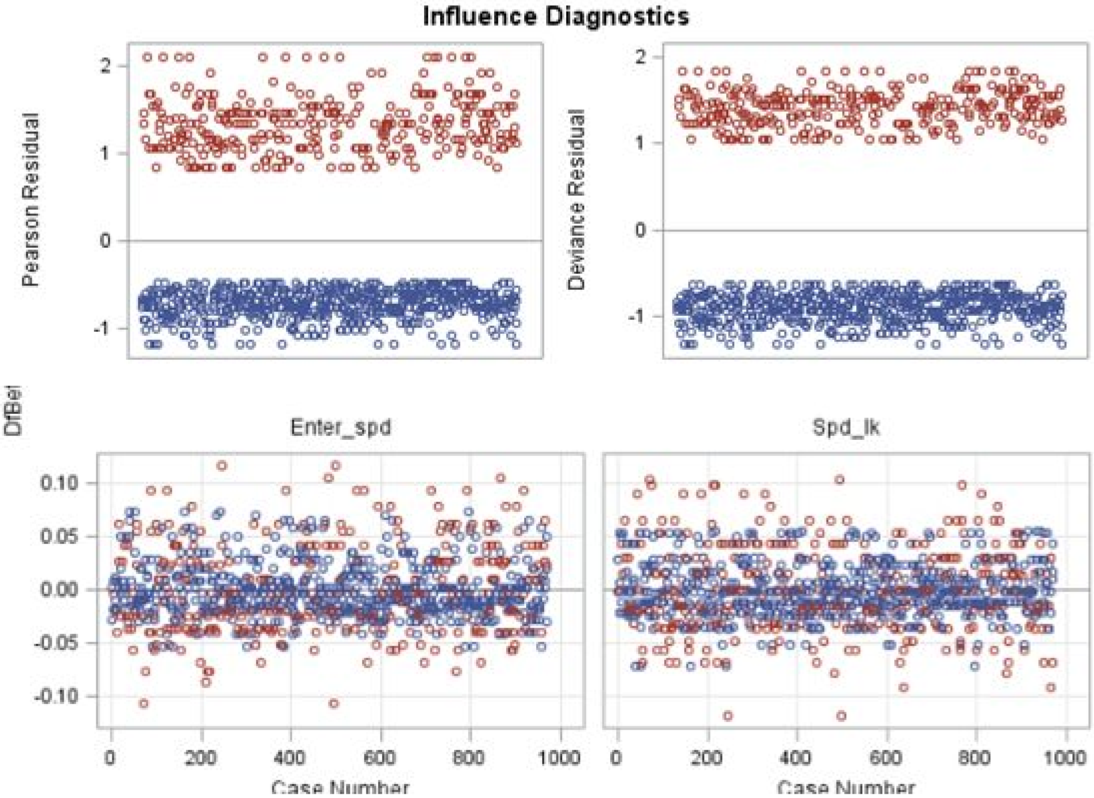
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Enter_spd	0.794	0.680	0.928
Spd_lk	0.790	0.677	0.921

- **Backward correlation table**

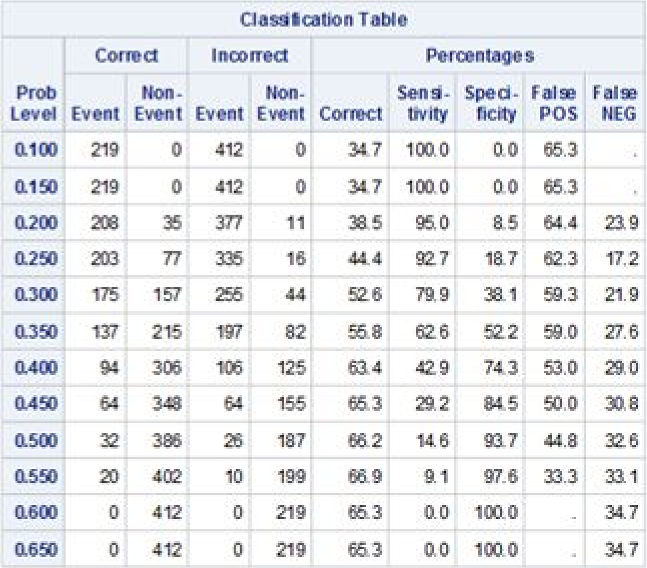


Estimated Correlation Matrix			
Parameter	Intercept	Enter_spd	Spd_lk
Intercept	1.0000	-0.5418	-0.4868
Enter_spd	-0.5418	1.0000	-0.4125
Spd_lk	-0.4868	-0.4125	1.0000

Backward diagnostics: outliers and influential points



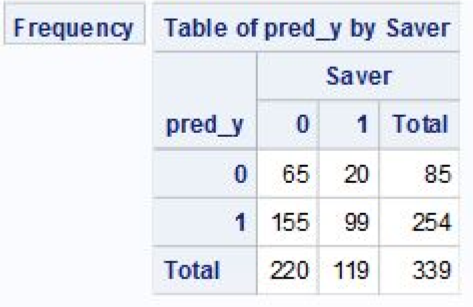
Classification table



Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.100	219	0	412	0	34.7	100.0	0.0	65.3	.
0.150	219	0	412	0	34.7	100.0	0.0	65.3	.
0.200	208	35	377	11	38.5	95.0	8.5	64.4	23.9
0.250	203	77	335	16	44.4	92.7	18.7	62.3	17.2
0.300	175	157	255	44	52.6	79.9	38.1	59.3	21.9
0.350	137	215	197	82	55.8	62.6	52.2	59.0	27.6
0.400	94	306	106	125	63.4	42.9	74.3	53.0	29.0
0.450	64	348	64	155	65.3	29.2	84.5	50.0	30.8
0.500	32	386	26	187	66.2	14.6	93.7	44.8	32.6
0.550	20	402	10	199	66.9	9.1	97.6	33.3	33.1
0.600	0	412	0	219	65.3	0.0	100.0	.	34.7
0.650	0	412	0	219	65.3	0.0	100.0	.	34.7

Classification metric



Frequency

Table of pred_y by Saver

	Saver		
pred_y	0	1	Total
0	65	20	85
1	155	99	254
Total	220	119	339

7. Code

Manish Singh section:

```
TITLE "Import Young peeople survey";
proc import datafiles="Young_People_Survey_Manish.xlsx"
out= responses_import dbms=xlsx replace;;
datarow=2;
getnames=yes;
run;
proc print data=responses_import;
run;

data responses_temporary original;
set responses_import;
array t(*) _numeric_;
do _n_=1 to dim(t);
if missing(t(_n_)) then flag=1;
end;
if not flag then output responses_temporary;
output original;
run;

proc print data = responses_temporary;
run;
/*create dummy variables*/
data responses_dummy;
set responses_temporary;
d_Gender= (Gender="male");
d_L_R_handed= (L_R_handed="right handed");
d_Only_child= (Only_child="yes");
d_Village_town= (Village_town="village");
d_House_flats= (House_flats="house/bungalow");
if Finances in (1,2,3) then d_Finances=0;
else if Finances in (4,5) then d_Finances=1;
if Education = "primary school" then d_Education=1;
else if Education = "secondary school" then d_Education=2;
else if Education = "college/bachelor degree" then d_Education=3;
else if Education = "masters degree" then d_Education=4;
else if Education = "currently a primary school pupil" then d_Education=5;
else if Education = "doctorate degree" then d_Education=6;
```

```
Run;
proc print data=responses_dummy;
run;

/*Full Model Diagnostics*/
proc logistic data=responses_dummy;
Model d_Finances(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating Age Siblings
d_Gender d_Only_child d_Village_town d_House_flats d_Education/stb;
run;

/*Full Model Diagnostics*/
proc logistic data=responses_dummy;
Model d_Finances(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating Age Siblings
d_Gender d_Only_child d_Village_town d_House_flats d_Education/corrb influence iplots;
run;

data responses_no_influence;
set responses_dummy;
if _n_=13 then delete;
if _n_=443 then delete;
if _n_=35 then delete;
if _n_=150 then delete;
if _n_=230 then delete;
if _n_=366 then delete;
if _n_=503 then delete;
if _n_=751 then delete;
if _n_=804 then delete;
if _n_=958 then delete;
if _n_=1007 then delete;
run;


proc logistic data=responses_no_influence;
Model d_Finances(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating Age Siblings
d_Gender d_Only_child d_Village_town d_House_flats d_Education/corrb influence iplots;
run;

*check interaction term;
proc logistic data=responses_no_influence;
```

```
model d_Finances(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating Age
        Siblings d_Gender d_Only_child d_Village_town d_House_flats d_Education
Age*Spending_on_healthy_eating Age*d_Education/ stb corrb;
run;

proc freq data=responses_no_influence;
tables d_Finances;
run;

proc surveyselect data=responses_no_influence out=train seed=495857
samprate=0.75 outall;
run;

proc print data=train;
run;

* check to see if the train/test split was done correctly;
proc freq data=train;
tables selected;
run;

data train;
set train;
if selected then train_y=d_Finances;
run;


/*Verifying full model*/
proc logistic data= train;
Model train_y(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating Age Siblings
d_Gender d_Only_child d_Village_town d_House_flats d_Education  /corrb;
run;


proc logistic data=train;
title "model selection forward";
model train_y(event='1')= Shopping_centres Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating Age Siblings
d_Gender d_Only_child d_Village_town d_House_flats d_Education /selection=forward rsquare
influence iplots corrb stb;
run;
```

```
proc logistic data=train;
title "model selection stepwise";
model train_y(event='1')= Shopping_centres Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating Age Siblings
d_Gender d_Only_child d_Village_town d_House_flats d_Education /selection=stepwise
rsquare influence iplots corrb stb;
run;

/*Check prediction data on full model*/
data newPrediction;
input Finances Shopping_centres     Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating
          Age     Siblings Gender $ Education $ Only_child $ Village_town $ House_flats $;

datalines;
. . . 1 1 . . . . . . . village .
. . . 5 5 . . . . . . . city .
;

data pred;
set newPrediction train;
d_Gender= (Gender="male");
d_L_R_handed= (L_R_handed="right handed");
d_Only_child= (Only_child="yes");
d_Village_town= (Village_town="village");
d_House_flats= (House_flats="house/bungalow");
if Finances in (1,2,3) then d_Finances=0;
else if Finances in (4,5) then d_Finances=1;
if Education = "primary school" then d_Education=1;
else if Education = "secondary school" then d_Education=2;
else if Education = "college/bachelor degree" then d_Education=3;
else if Education = "masters degree" then d_Education=4;
else if Education = "currently a primary school pupil" then d_Education=5;
else if Education = "doctorate degree" then d_Education=6;
run;
proc print;
run;
* logistic regression model;
proc logistic data=pred;
model d_Finances(event='1')= Entertainment_spending Spending_on_looks d_Village_town ;
output out=pred p=phat lower=lcl upper=ucl predprob=(individual);
run;
```

```
proc print;
run;

/*Prediction Model ends*/

* Find the final model, compute predicted value on training set, obtain the cut-off value for p;
proc logistic data=train;
title "find cut-off";
model train_y(event='1')= Entertainment_spending Spending_on_looks d_Village_town /ctable
pprob= (0.1 to 0.6 by 0.05);
*save predictions in sas dataset "pred";
output out=pred(where=(train_y=.))  p=phat lower=lcl upper=ucl
           predprob=(individual);
run;

proc print data=pred;
run;

/*Using cut-off value to compute classification matrix*/
data probs;
set pred;
pred_y = 0;
if (phat > 0.35) then pred_y = 1;
run;

proc print data=probs;
run;

* compute classification matrix;
proc freq data=probs;
tables d_Finances*pred_y/norow nocol nopercent;
run;

/*making a training set of 60 START*/
proc surveyselect data=responses_no_influence out=train60set seed=15865
samprate=0.60 outall;
run;
proc print data=train60set;
run;

* check to see if the train/test split was done correctly;
proc freq data=train60set;
tables selected;
```

```
run;

data train60set;
set train60set;
if selected then train_y=d_Finances;
run;

proc logistic data=train60set;
title "model selection forward";
model train_y(event='1')= Shopping_centres Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating Age Siblings
d_Gender d_Only_child d_Village_town d_House_flats d_Education /selection=forward rsquare
influence iplots corrb stb;
run;


* Find the final model, compute predicted value on training set, obtain the cut-off value for p;
proc logistic data=train60set;
title "find cut-off";
model train_y(event='1')= Entertainment_spending Spending_on_looks d_Village_town /ctable
pprob= (0.1 to 0.6 by 0.05);
*save predictions in sas dataset "pred";
output out=pred(where=(train_y=.))  p=phat lower=lcl upper=ucl
          predprob=(individual);
run;

proc print data=pred;
run;

/*Using cut-off value to compute classification matrix*/
data probs;
set pred;
pred_y = 0;
if (phat > 0.3) then pred_y = 1;
run;

proc print data=probs;
run;

* compute classification matrix;
proc freq data=probs;
tables d_Finances*pred_y/norow nocol nopercent;
run;
```

Michal Chowaniak:

```
*Michal Chowaniak Project model B;

/* PLEASE READ THIS FIRST BEFORE EXECUTING THIS FILE

There are two data files needed to execute this SAS file.

1. Michal_Chowaniak_project_dataset.csv
2. Michal_Chowaniak_project_dataset_predictions.csv

There are two infile statements which have to be updated if this SAS file is executed in differEnt
folder than below.

1. First infile statement is on line 20
2. Second infile statement is on line 232

*/

*import data set;
data youngpeople;
infile "S:\CSC423\Project\Michal_Chowaniak_project_dataset.csv" firstobs=2 delimiter=','
missover;
input finances shopping clothing entertainment looks gadgets eating age height weight siblings
gender $ l_r_handed $ education $ only_child $ village_town $ house_apt $;
run;

title "young people dataset";
proc print;
run;

*Create boxplot for finances  by age;
TITLE "Boxplots - finance by age";
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT age*finances ;
RUN;
```

```
*collapse variables to 0-no, 1=yes variable;
data youngpeople;
set youngpeople;

*I save all the money I can: 0-no, 1-yes;
if finances = 1 then b_fin = 0;
if finances = 2 then b_fin = 0;
if finances = 3 then b_fin = 0;
if finances = 4 then b_fin = 1;
if finances = 5 then b_fin = 1;

*I enjoy going to large shopping centres: 0 no, 1 yes;
if shopping = 1 then b_shop = 0;
if shopping = 2 then b_shop = 0;
if shopping = 3 then b_shop = 0;
if shopping = 4 then b_shop = 1;
if shopping = 5 then b_shop = 1;

*I prefer branded clothing to non branded: 0-no, 1-yes;
if clothing = 1 then b_clo = 0;
if clothing = 2 then b_clo = 0;
if clothing = 3 then b_clo = 0;
if clothing = 4 then b_clo = 1;
if clothing = 5 then b_clo = 1;

*I spend a lot of money on partying and socializing: 0-no, 1-yes;
if entertainment = 1 then b_ent = 0;
if entertainment = 2 then b_ent = 0;
if entertainment = 3 then b_ent = 0;
if entertainment = 4 then b_ent = 1;
if entertainment = 5 then b_ent = 1;

*I spend a lot of money on my appearance: 0-no, 1-yes;
if looks = 1 then b_look = 0;
if looks = 2 then b_look = 0;
if looks = 3 then b_look = 0;
if looks = 4 then b_look = 1;
if looks = 5 then b_look = 1;

*I spend a lot of money on gadgets: 0-no, 1-yes;
if gadgets = 1 then b_gadg = 0;
if gadgets = 2 then b_gadg = 0;
if gadgets = 3 then b_gadg = 0;
if gadgets = 4 then b_gadg = 1;
if gadgets = 5 then b_gadg = 1;

*I will hapilly pay more money for good, quality or healthy food: 0-no, 1-yes;
if eating = 1 then b_eat = 0;
if eating = 2 then b_eat = 0;
if eating = 3 then b_eat = 0;
if eating = 4 then b_eat = 1;
if eating = 5 then b_eat = 1;

*Gender: 0-female, 1-male;
if gender = 'female' then b_gender = 0;
if gender = 'male' then b_gender = 1;

*I am: Left handed - Right handed: 0-left, 1-right;
if l_r_handed = 'left han' then b_hand = 0;
if l_r_handed = 'right ha' then b_hand = 1;

*Highest education achieved: Currently a Primary school pupil, Primary school, Secondary
school, College/Bachelor degree, masters degree, doctorate degree;
* bachelor and below 0 , above bachelor = 1;
if education = 'secondar' then b_edu = 0;
if education = 'college/' then b_edu = 0;
if education = 'primary' then b_edu = 0;
if education = 'currentl' then b_edu = 0;
if education = 'doctorat' then b_edu = 1;
if education = 'masters' then b_edu = 1;

*I am the only child: : 0-no, 1-yes;
if only_child = 'no' then b_child = 0;
if only_child = 'yes' then b_child = 1;

*I spent most of my childhood in a: City - 0, village - 1;
if village_town = 'city' then b_town = 0;
if village_town = 'village' then b_town = 1;

*I lived most of my childhood in a: house/bungalow - 1, block of flats - 0;
if house_apt = 'block of' then b_house = 0;
if house_apt = 'house/bu' then b_house = 1;

title "young people dataset";
proc print;
```

```
run;

*Create boxplot for finances  by age;
TITLE "Boxplots - finance by age";
PROC SORT;
BY b_fin;
RUN;
PROC BOXPLOT;
PLOT age*b_fin ;
RUN;

*Create frequency tables for y variable;
title "Frequency";
proc freq;
tables b_fin;
run;

*Check - Multicollinearity, outliers and influencial points;
title Multicollinearity, outliers and influencial points;
proc logistic;
model b_fin ( event='1') = b_shop b_clo b_ent b_look b_gadg b_eat b_gender b_hand b_edu
b_child b_town b_house age height weight siblings weight*height/ corrb influence iplots;
run;

*Remove height, correclation over 0.9;
data youngpeople;
set youngpeople;
drop height;
drop weight*height;

*Check - Multicollinearity, outliers and influencial points;
title Multicollinearity, outliers and influencial points;
proc logistic;
model b_fin ( event='1') = b_shop b_clo b_ent b_look b_gadg b_eat b_gender b_hand b_edu
b_child b_town b_house age weight siblings / corrb influence iplots stb;
run;

*no outliers;

*some ifluetial points - keep them;

*full model;
```

```
/* Log( b_fin = 1/ b_fin =0 ) = -0.3163 + 0.0127*b_shop + 0.1247*b_clo – 0.5820*b_ent –
0.5332*b_look +
0.0937*b_gadg + 0.3124*b_eat + 0.0454*b_gender – 0.1129*b_hand + 0.00638*b_edu +
0.1186*b_child +
0.4420*b_town – 0.0106*b_house + 0.0332*age – 0.0126* weight – 0.0541*siblings
*/

*Split data into training and test sets;
title "Training and Test sets";
proc surveyselect data=youngpeople out=youngpeople_train8020 seed=831957
samprate=0.7 outall;
run;

proc print data=youngpeople_train8020;
run;

* check to see if the train/test split was done correctly;
title "Training and Test sets frequency";
proc freq data=youngpeople_train8020;
tables selected;
run;

*create new y variable  for training set, and = NA for testing set;
data youngpeople_train8020;
set youngpeople_train8020;
*if selected is equal to 1 then;
if selected = 1 then b_fin_train = b_fin; *y variable for training data set;
run;

title "Training set";
proc print;
run;

* run selection backward method on training set;
title Training set model selection - Backward;
proc logistic data=youngpeople_train8020;
model b_fin_train (event='1')= b_shop b_clo b_ent b_look b_gadg b_eat b_gender b_hand
b_edu b_child b_town b_house age weight siblings /
  selection=backward rsquare;
run;

* run selection stepwise method on training set;
title Training set model selection - Stepwise;
```

```
proc logistic data=youngpeople_train8020;
model b_fin_train (event='1')= b_shop b_clo b_ent b_look b_gadg b_eat b_gender b_hand
b_edu b_child b_town b_house age weight siblings /
  selection=stepwise rsquare;
run;

*Final Model;
*Log(p/(1-p) = -0.5925 – 0.5726*b_ent – 0.6075*b_look + 0.4138*b_eat + 0.4318*b_town;

*Double Check - Multicollinearity, outliers and influencial points - final model;
title Multicollinearity, outliers and influencial points;
proc logistic data=youngpeople_train8020; ;
model b_fin_train ( event='1') = b_ent b_look b_eat b_town / corrb influence iplots stb;
run;

/*
**************I WAS UNABLE TO USE DATALINES TO INPUT DATA WITH A SPACE
BETWEEN, SO I DECIDED TO INPORT DATA FROM A FILE**************

*Compute predictions for a person who  spends money on entertainment
            ;
data new;
input finances shopping clothing entertainment looks gadgets eating age height weight siblings
gender $ l_r_handed $ 8 education $ 8 only_child $ village_town $ 8 house_apt $ 8;
datalines;
. 4 4 5 4 2 4 20 150 59 2 male right ha college/ no    village    block of
;
proc print data=new;
run;
*/

**************I WAS UNABLE TO USE DATALINES TO INPUT DATA WITH A SPACE
BETWEEN, SO I DECIDED TO INPORT DATA FROM A FILE**************;
*Compute predictions for a person who  spends money on entertainment
*import data set;
data new;
infile "S:\CSC423\Project\Michal_Chowaniak_project_dataset_predictions.csv" firstobs=2
delimiter=',' missover;
input finances shopping clothing entertainment looks gadgets eating age height weight siblings
gender $ l_r_handed $ education $ only_child $ village_town $ house_apt $;
run;

proc print data=new;
```

```
run;

*collapse variables to 0-no, 1=yes variable;
data pred;
set new youngpeople;

*I save all the money I can: 0-no, 1-yes;
if finances = 1 then b_fin = 0;
if finances = 2 then b_fin = 0;
if finances = 3 then b_fin = 0;
if finances = 4 then b_fin = 1;
if finances = 5 then b_fin = 1;

*I enjoy going to large shopping centres: 0 no, 1 yes;
if shopping = 1 then b_shop = 0;
if shopping = 2 then b_shop = 0;
if shopping = 3 then b_shop = 0;
if shopping = 4 then b_shop = 1;
if shopping = 5 then b_shop = 1;

*I prefer branded clothing to non branded: 0-no, 1-yes;
if clothing = 1 then b_clo = 0;
if clothing = 2 then b_clo = 0;
if clothing = 3 then b_clo = 0;
if clothing = 4 then b_clo = 1;
if clothing = 5 then b_clo = 1;

*I spend a lot of money on partying and socializing: 0-no, 1-yes;
if entertainment = 1 then b_ent = 0;
if entertainment = 2 then b_ent = 0;
if entertainment = 3 then b_ent = 0;
if entertainment = 4 then b_ent = 1;
if entertainment = 5 then b_ent = 1;

*I spend a lot of money on my appearance: 0-no, 1-yes;
if looks = 1 then b_look = 0;
if looks = 2 then b_look = 0;
if looks = 3 then b_look = 0;
if looks = 4 then b_look = 1;
if looks = 5 then b_look = 1;

*I spend a lot of money on gadgets: 0-no, 1-yes;
if gadgets = 1 then b_gadg = 0;
if gadgets = 2 then b_gadg = 0;
if gadgets = 3 then b_gadg = 0;
if gadgets = 4 then b_gadg = 1;
if gadgets = 5 then b_gadg = 1;

*I will hapilly pay more money for good, quality or healthy food: 0-no, 1-yes;
if eating = 1 then b_eat = 0;
if eating = 2 then b_eat = 0;
if eating = 3 then b_eat = 0;
if eating = 4 then b_eat = 1;
if eating = 5 then b_eat = 1;

*Gender: 0-female, 1-male;
if gender = 'female' then b_gender = 0;
if gender = 'male' then b_gender = 1;

*I am: Left handed - Right handed: 0-left, 1-right;
if l_r_handed = 'left han' then b_hand = 0;
if l_r_handed = 'right ha' then b_hand = 1;

*Highest education achieved: Currently a Primary school pupil, Primary school, Secondary
school, College/Bachelor degree, masters degree, doctorate degree;
* bachelor and below 0 , above bachelor = 1;
if education = 'secondar' then b_edu = 0;
if education = 'college/' then b_edu = 0;
if education = 'primary' then b_edu = 0;
if education = 'currentl' then b_edu = 0;
if education = 'doctorat' then b_edu = 1;
if education = 'masters' then b_edu = 1;

*I am the only child: : 0-no, 1-yes;
if only_child = 'no' then b_child = 0;
if only_child = 'yes' then b_child = 1;

*I spent most of my childhood in a: City - 0, village - 1;
if village_town = 'city' then b_town = 0;
if village_town = 'village' then b_town = 1;

*I lived most of my childhood in a: house/bungalow - 1, block of flats - 0;
if house_apt = 'block of' then b_house = 0;
if house_apt = 'house/bu' then b_house = 1;
run;
```

```
* logistic regression model;
title "logistic regression model";
proc logistic data=pred;
model b_fin ( event='1') = b_ent b_look b_eat b_town;
output out=pred p=phat lower=lcl upper=ucl;
run;

*printing predicted probabilities and confidence intervals;
proc print data=pred;
title 'Predicted Probabilities and 95% Confidence Limits';
run;

*Generate classification table and identify cutoff value;
title "Cut off value";
proc logistic data=youngpeople_train8020; ;
model b_fin_train ( event='1') = b_ent b_look b_eat b_town / ctable pprob =(0.1 to 0.6 by 0.05);

*compute predicted probability for test set;
output out=youngpeople_test8020(where=(b_fin_train=.)) p=phat lower = lcl upper =ucl;
run;

*print output table for test set predicted prob.;
title "Output table for test set predicted probability";
proc print data = youngpeople_test8020;
run;

*use cut-off value to compute classification matrics, create new data set for probabilities;
data youngpeople_prob8020;
set youngpeople_test8020;
b_fin_prob = 0;
if (phat > 0.40) then b_fin_prob = 1;
run;

proc print data = youngpeople_prob8020;
run;

*Classification matrix. table observed y b_fin and predicted y b_fin_prob;
title 'observed y and predicted y';
proc freq data = youngpeople_prob8020;
tables b_fin * b_fin_prob / norow nocol nopercent;
run;
```

Priyank Beno Cerejo section:

```
/*Import Data*/
TITLE "Import Young peeople survey";
proc import datafiles="C:\Users\pcerejo\Downloads\Priyank_responses.xlsx"
out= responses dbms=xlsx replace;;
datarow=2;
getnames=yes;
run;
proc print data=responses;
run;

proc contents varnum;
run;


/* Deleting the Missing Value */
data need original;
set responses;
array t(*) _numeric_;
do _n_=1 to dim(t);
if missing(t(_n_)) then flag=1;
end;
if not flag then output need;
output original;
run;

/*create dummy variables*/
/* Imputing(0,1) for finances where 0 is(1,2,3) and 1 is(4,5) */
/* Where 1,2,3 are Not money savers(Spenders)
4,5 are money savers*/

data temp;
set need;
array t(*) Finances;
do _n_=1 to dim(t);
if t(_n_) in (1,2,3) then t(_n_)=0;
else if t(_n_) in (4,5) then t(_n_)=1;
end;
New_Gender= (Gender="male");
New_L_R_handed= (L_R_handed="right handed");
```

```
New_Only_child= (Only_child="yes");
New_Village_town= (Village_town="village");
New_House_flats= (House_flats="house/bungalow");
if Education = "primary school" then New_Education=1;
else if Education = "secondary school" then New_Education=2;
else if Education = "college/bachelor degree" then New_Education=3;
else if Education = "masters degree" then New_Education=4;
else if Education = "currently a primary school pupil" then New_Education=5;
else if Education = "doctorate degree" then New_Education=6;
Run;
proc print data=temp;
run;


proc freq data=temp;
tables _numeric_;
run;

/*To check the most infulential predictor*/
proc logistic data=temp;
Model Finances(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating Age New_Gender
New_L_R_handed New_Only_child New_Village_town New_House_flats New_Education/stb;
run;

/*To check interaction term*/
proc logistic data=temp;
Model Finances(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Age*Spending_on_healthy_eating Age
New_Gender New_L_R_handed New_Only_child Age*New_Village_town New_House_flats
New_Education/stb corrb;
run;


/*Full Model Diagnostics*/
/*To check Std coefficient and R-squared*/
ods graphics on;
proc logistic data=temp;
Model Finances(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating Age New_Gender
New_L_R_handed New_Only_child New_Village_town New_House_flats New_Education/stb
corrb influence iplots;
run;
```

```
ods graphics off;


* Split the data into training and test sets - 75/25;
* samprate = 75% of observations to be randomly selected for training set
* out = train defines new sas dataset for training/test sets;

proc surveyselect data=temp out=train seed=495857
samprate=0.75 outall;
run;

proc print data=train;
run;

* check to see if the train/test split was done correctly;
proc freq data=train;
tables selected;
run;


*create new variable new_y = Finances for training set, and = NA for testing set;
data train;
set train;
if selected then train_y=Finances;
run;


/*Stepwise Selection*/
proc logistic data=train;
Model train_y(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating Age New_Gender
New_L_R_handed New_Only_child New_Village_town New_House_flats
New_Education/selection=stepwise rsquare;
run;


/*Backward Selection*/
proc logistic data=train;
Model train_y(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating Age New_Gender
New_L_R_handed New_Only_child New_Village_town New_House_flats
New_Education/selection=Backward rsquare;
run;
```

```
/*Check prediction data on full model*/
data CheckPred;
input Finances Shopping_centres     Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating
        Age     Siblings Gender $ Education $ Only_child $ Village_town $ House_flats $;

datalines;
. . . 1 2 . . . . . . . village .
. . . 5 4 . . . . . . . city .
;

data pred;
set CheckPred train;
array t(*) Finances;
do _n_=1 to dim(t);
if t(_n_) in (1,2,3) then t(_n_)=0;
else if t(_n_) in (4,5) then t(_n_)=1;
end;
New_Gender= (Gender="male");
New_L_R_handed= (L_R_handed="right handed");
New_Only_child= (Only_child="yes");
New_Village_town= (Village_town="village");
New_House_flats= (House_flats="house/bungalow");
if Education = "primary school" then New_Education=1;
else if Education = "secondary school" then New_Education=2;
else if Education = "college/bachelor degree" then New_Education=3;
else if Education = "masters degree" then New_Education=4;
else if Education = "currently a primary school pupil" then New_Education=5;
else if Education = "doctorate degree" then New_Education=6;
Run;
proc print;
run;
* logistic regression model;
proc logistic data=pred;
model train_y(event='1')= Entertainment_spending Spending_on_looks New_Village_town ;
output out=pred p=phat lower=lcl upper=ucl predprob=(individual);
run;
proc print data=pred;
run;
```

```
* Find the final model, compute predicted value on training set, obtain the cut-off value for p;
proc logistic data=train;
model train_y(event='1')= Entertainment_spending Spending_on_looks New_Village_town
/ctable pprob= (0.1 to 0.6 by 0.05);
*save predictions in sas dataset "pred";
output out=pred(where=(train_y=.))  p=phat lower=lcl upper=ucl
            predprob=(individual);
run;

proc print data=pred;
run;


/*USing cut-off value to compute classification matrix*/
data probs;
set pred;
pred_y = 0;
if (phat > 0.4) then pred_y = 1;
run;

proc print data=probs;
run;

* compute classification matrix;
proc freq data=probs;
tables finances*pred_y/norow nocol nopercent;
run;

-----------------------------------------------------------------
*2nd Model;

* Split the data into training and test sets - 60/40;
* samprate = 60% of observations to be randomly selected for training set
* out = train defines new sas dataset for training/test sets;

proc surveyselect data=temp out=train60 seed=587634
samprate=0.60 outall;
run;

proc print data=train60;
run;

* check to see if the train60/test split was done correctly;
```

```
proc freq data=train60;
tables selected;
run;


*create new variable new_y = Finances for training set, and = NA for testing set;
data train60;
set train60;
if selected then train_y=Finances;
run;


/*Stepwise Selection*/
proc logistic data=train60;
Model train_y(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating Age New_Gender
New_L_R_handed New_Only_child New_Village_town New_House_flats
New_Education/selection=stepwise rsquare;
run;


/*Backward Selection*/
proc logistic data=train60;
Model train_y(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending_on_looks Spending_on_gadgets Spending_on_healthy_eating Age New_Gender
New_L_R_handed New_Only_child New_Village_town New_House_flats
New_Education/selection=Backward rsquare;
run;


* Find the final model, compute predicted value on training set, obtain the cut-off value for p;
proc logistic data=train60;
model train_y(event='1')= Entertainment_spending Spending_on_looks New_Village_town
/ctable pprob= (0.1 to 0.6 by 0.05);
*save predictions in sas dataset "pred";
output out=pred(where=(train_y=.))  p=phat lower=lcl upper=ucl
           predprob=(individual);
run;

proc print data=pred;
run;
```

```
/*USing cut-off value to compute classification matrix*/
data probs;
set pred;
pred_y = 0;
if (phat > 0.35) then pred_y = 1;
run;

proc print data=probs;
run;

* compute classification matrix;
proc freq data=probs;
tables finances*pred_y/norow nocol nopercent;
run;
```

Rushabh Shah section:

```
TITLE 'Responses - Import';
PROC IMPORT datafile="responses.csv" out=responses replace;
delimiter=',';
getnames=YES;
datarow=2;
RUN;
proc print;
run;
data responses;
set responses;
if Smoking='never smoked' then Smoke=0;
        else if Smoking='tried smoking' then Smoke=1;
        else if Smoking='former smoker' then Smoke=2;
        else Smoke=3;
if Alcohol='never' then Drink=0;
        else if Alcohol='social drinker' then Drink=1;
        else Drink=2;
if Punctuality='i am always on time' then Punctual=0;
        else if Punctuality='i am often early' then Punctual=1;
        else Punctual=2;
if Lying='never' then Lie=0;
        else if Lying='only to avoid hurting someone' then Lie=1;
        else if Lying='sometimes' then Lie=2;
        else Lie=3;
if Internet_usage='no time at all' then Internet_use=0;
        else if Internet_usage='less than an hour a day' then Internet_use=1;
        else if Internet_usage='few hours a day' then Internet_use=2;
        else Internet_use=3;
if Gender='female' then sex=0;
        else sex=1;
if Left___right_handed='right handed' then Handed=0;
        else Handed=1;
if Education='currently a primary schoo' then schooling=0;
        else if Education='primary school' then schooling=1;
        else if Education='secondary school' then schooling=2;
        else schooling=3;
if Only_child='no' then Only=0;
        else Only=1;
if Village___town='village' then Urban=0;
        else Urban=1;
if House___block_of_flats='block of flats' then Housing=0;
        else Housing=1;
if Finances=1 then finances=0;
        else if Finances=2 then finances=0;
        else if Finances=3 then finances=0;;
        else finances=1;
Run;
Title “Housing vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT Housing*finances;
RUN;
Title “Urban vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT Urban*finances;
RUN;
Title “Schooling vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
```

```
PLOT schooling*finances;
RUN;
Title “Only Child status vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT Only*finances;
RUN;
Title “Left or right handed vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT Handed*finances;
RUN;
Title “Gender vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT sex*finances;
RUN;
Title “Number of Siblings vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT Number_of_siblings*finances;
RUN;
Title “Weight vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT Weight*finances;
RUN;
Title “Height vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT Height*finances;
```

```
RUN;
Title “Age vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT Age*finances;
RUN;
Title “Healthy food spending vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT Spending_on_healthy_eating*finances;
RUN;
Title “Spending on gadgets vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT Spending_on_gadgets*finances;
RUN;
Title “Spending on looks vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT Spending_on_looks*finances;
RUN;
Title “Entertainment spending vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT Entertainment_spending*finances;
RUN;
Title “Branded clothing vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT Branded_clothing*finances;
RUN;
```

```
Title “Shopping Centers vs Finances”;
PROC SORT;
BY finances;
RUN;
PROC BOXPLOT;
PLOT Shopping_centres*finances;
RUN;
Title "Full Logistic Regression Model";
proc logistic data=responses;
model Finances = Housing Urban schooling Only Handed sex Number_of_siblings Weight
Height Age Spending_on_healthy_eating Spending_on_gadgets Spending_on_looks
Entertainment_spending Branded_clothing Shopping_centres;
RUN;
data new;
input Housing Urban schooling Only Handed sex Number_of_siblings Weight Height Age
Spending_on_healthy_eating Spending_on_gadgets Spending_on_looks
Entertainment_spending Branded_clothing Shopping_centres
datalines;
0 1 2 0 0 1 1 77 186 20 4 4 1 3 3 3
1 3 0 0 0 2 58 163 19 2 5 2 4 1 4
;
data new;
set new responses;
run;
proc print data=new;
run;
proc logistic data=new;
model finances(event='1')=Housing Urban schooling Only Handed sex Number_of_siblings
Weight Height Age Spending_on_healthy_eating Spending_on_gadgets Spending_on_looks
Entertainment_spending Branded_clothing Shopping_centres;
output out=fullpred p=phat lower= lcl upper=ucl;
run;
proc print;
run;
proc surveyselect data=responses samprate=0.60 seed=12345 out=train outall
          method=srs noprint;
        run;
Title "Full Logistic Regression Model";
proc logistic data=train;
model finances = Housing Urban schooling Only Handed sex Number_of_siblings Weight
Height Age Spending_on_healthy_eating Spending_on_gadgets Spending_on_looks
Entertainment_spending Branded_clothing Shopping_centre
```

```
/selection=forward;
TITLE"Final Model";
proc logistic DATA=train;
model finances = Urban Entertainment_spending
/corrb influence iplots;
Run;
data new1;
Input Urban Entertainment_spending;
1 2
0 1
;
data new1;
set new1 responses;
run;
proc print data=new1;
run;
proc logistic data=new1;
model finances(event='1')= Urban Entertainment_spending;
output out=fpred p=phat lower= lcl upper=ucl;
run;
proc print;
run;
```

Valentine Silvester Correia section:

```
TITLE "Import Young peeople survey";
proc import datafiles="C:\Users\VCORREIA\Desktop\Final Project
New\young_people_survey_valentine.xlsx"
out= responses_import dbms=xlsx replace;
datarow=2;
getnames=yes;
run;
proc print data=responses_import;
run;

data responses_temp original;
set responses_import;
array t(*) _numeric_;
do _n_=1 to dim(t);
if missing(t(_n_)) then flag=1;
end;
if not flag then output responses_temp;
output original;
run;

/*Find frequency*/
proc freq data=responses_temp;
tables _numeric_;
run;

/*create dummy variables*/
data d_responses;
set responses_temp;
d_Gender= (Gender="male");
d_L_R_handed= (L_R_handed="right handed");
d_Only_child= (Only_child="yes");
d_Village_town= (Village_town="village");
d_House_flats= (House_flats="house/bungalow");
if Finances in (1,2,3) then d_Finances=0;
else if Finances in (4,5) then d_Finances=1;
if Education = "primary school" then d_Education=1;
else if Education = "secondary school" then d_Education=2;
else if Education = "college/bachelor degree" then d_Education=3;
else if Education = "masters degree" then d_Education=4;
else if Education = "currently a primary school pupil" then d_Education=5;
else if Education = "doctorate degree" then d_Education=6;
Run;
proc print data=d_responses;
run;

/*frequency savers*/
proc freq data= d_responses;
tables d_Finances;
run;

/*Boxplot- Finances before binning*/
PROC SORT;
BY Finances;
RUN;
PROC BOXPLOT;
PLOT Age*Finances;
RUN;

/*Boxplot- Finances after binning*/
PROC SORT;
BY d_Finances;
RUN;
PROC BOXPLOT;
PLOT Age*d_Finances;
RUN;

/*Creation of Histogram*/
proc univariate data= d_responses;
var Finances;
histogram / normal (mu=est sigma=est);
run;

/*Checking interaction terms*/
proc logistic data=d_responses;
model d_Finances(evet='1') = Branded_clothing*Spending__on_looks Age*d_Education/ stb
corrb;
run;


/*Verifying full model*/
proc logistic data= d_responses;
```

```
Model d_Finances(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending__on_looks Spending_on_gadgets Spending_on_healthy_eating
Age Siblings d_Gender d_Only_child d_Village_town d_House_flats d_Education  / stb;
run;

/*Verifying full model diagnostics*/
proc logistic data=d_responses;
Model d_Finances(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending__on_looks Spending_on_gadgets Spending_on_healthy_eating
Age Siblings d_Gender d_Only_child d_Village_town d_House_flats d_Education / corrb
influence iplots;
run;

data r_no_influence;
set d_responses;
if _n_=13 then delete;
if _n_=443 then delete;
if _n_=35 then delete;
if _n_=150 then delete;
if _n_=230 then delete;
if _n_=366 then delete;
if _n_=503 then delete;
if _n_=751 then delete;
if _n_=804 then delete;
if _n_=958 then delete;
if _n_=1007 then delete;
run;

proc logistic data=r_no_influence;
Model d_Finances(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending__on_looks Spending_on_gadgets Spending_on_healthy_eating
Age Siblings d_Gender d_Only_child d_Village_town d_House_flats d_Education / corrb
influence iplots;
run;

proc freq data=r_no_influence;
tables d_Finances;
run;

proc surveyselect data=r_no_influence out=train seed=564786
samprate=0.8 outall;
run;
```

```
proc print data=train;
run;

* check to see if the train/test split was done correctly;
proc freq data=train;
tables selected;
run;


data train;
set train;
if selected then train_y=d_Finances;
run;

/*Verifying full model*/
proc logistic data= train;
Model train_y(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending__on_looks
Spending_on_gadgets Spending_on_healthy_eating Age Siblings d_Gender d_Only_child
d_Village_town d_House_flats d_Education  / corrb;
run;

/* model selection stepwise */
proc logistic data=train;
title "model selection stepwise";
model d_Finances(event='1')= Shopping_centres Branded_clothing Entertainment_spending
Spending__on_looks Spending_on_gadgets Spending_on_healthy_eating
Age Siblings d_Gender d_Only_child d_Village_town d_House_flats
d_Education /selection=stepwise rsquare influence iplots corrb stb;
run;

/* model selection forward */
proc logistic data=train;
title "model selection forward";
model d_Finances(event='1')= Shopping_centres Branded_clothing Entertainment_spending
Spending__on_looks Spending_on_gadgets Spending_on_healthy_eating
Age Siblings d_Gender d_Only_child d_Village_town d_House_flats
d_Education /selection=forward rsquare influence iplots corrb stb;
run;


/*Check prediction data on full model*/
data newPrediction;
```

```
input Finances Shopping_centres            Branded_clothing Entertainment_spending
Spending__on_looks Spending_on_gadgets Spending_on_healthy_eating
        Age    Siblings Gender $ Education $ Only_child $ Village_town $ House_flats $;

datalines;
. . . 1 2 . . . . . . . village .
. . . 4 5 . . . . . . . city .
;

data pred;
set newPrediction train;
d_Gender= (Gender="male");
d_L_R_handed= (L_R_handed="right handed");
d_Only_child= (Only_child="yes");
d_Village_town= (Village_town="village");
d_House_flats= (House_flats="house/bungalow");
if Finances in (1,2,3) then d_Finances=0;
else if Finances in (4,5) then d_Finances=1;
if Education = "primary school" then d_Education=1;
else if Education = "secondary school" then d_Education=2;
else if Education = "college/bachelor degree" then d_Education=3;
else if Education = "masters degree" then d_Education=4;
run;
proc print;
run;

* logistic regression model;
proc logistic data=pred;
model d_Finances(event='1')= Entertainment_spending Spending__on_looks d_Village_town;
output out=pred p=phat lower=lcl upper=ucl predprob=(individual);
run;
proc print;
run;


* Find the final model, compute predicted value on training set, obtain the cut-off value for p;
proc logistic data=train;
title "find cut-off";
model train_y(event='1')= Entertainment_spending Spending__on_looks d_Village_town /ctable
pprob= (0.1 to 0.6 by 0.05);
*save predictions in sas dataset "pred";
output out=pred(where=(train_y=.))  p=phat lower=lcl upper=ucl
        predprob=(individual);
```

```
run;

proc print data=pred;
run;

/*Using cut-off value to compute classification matrix*/
data probs;
set pred;
pred_y = 0;
if (phat > 0.35) then pred_y = 1;
run;

proc print data=probs;
run;

* compute classification matrix;
proc freq data=probs;
tables d_Finances*pred_y/norow nocol nopercent;
run;


/*second model with training:testing data 70:30*/
proc surveyselect data=r_no_influence out=train1 seed=475684
samprate=0.7 outall;
run;

proc print data=train1;
run;

* check to see if the train/test split was done correctly;
proc freq data=train1;
tables selected;
run;

data train1;
set train1;
if selected then train_y1=d_Finances;
run;

/*Verifying full model*/
proc logistic data= train1;
Model train_y1(event='1') = Shopping_centres Branded_clothing Entertainment_spending
Spending__on_looks
```

```
Spending_on_gadgets Spending_on_healthy_eating Age Siblings d_Gender d_Only_child
d_Village_town d_House_flats d_Education  / corrb;
run;

/* model selection backward */
proc logistic data=train1;
title "model selection backward";
model d_Finances(event='1')= Shopping_centres Branded_clothing Entertainment_spending
Spending__on_looks Spending_on_gadgets Spending_on_healthy_eating
Age Siblings d_Gender d_Only_child d_Village_town d_House_flats
d_Education /selection=backward rsquare influence iplots corrb stb;
run;

/* model selection stepwise */
proc logistic data=train1;
title "model selection stepwise";
model d_Finances(event='1')= Shopping_centres Branded_clothing Entertainment_spending
Spending__on_looks Spending_on_gadgets Spending_on_healthy_eating
Age Siblings d_Gender d_Only_child d_Village_town d_House_flats
d_Education /selection=stepwise rsquare influence iplots corrb stb;
run;


/*Check prediction data on full model*/
data newPrediction1;
input Finances Shopping_centres              Branded_clothing Entertainment_spending
Spending__on_looks Spending_on_gadgets Spending_on_healthy_eating
        Age     Siblings Gender $ Education $ Only_child $ Village_town $ House_flats $;

datalines;
. . . 1 1 . . . . . . . city .
. . . 4 4 . . . . . . . village .
;

data pred1;
set newPrediction1 train1;
d_Gender= (Gender="male");
d_L_R_handed= (L_R_handed="right handed");
d_Only_child= (Only_child="yes");
d_Village_town= (Village_town="village");
d_House_flats= (House_flats="house/bungalow");
if Finances in (1,2,3) then d_Finances=0;
else if Finances in (4,5) then d_Finances=1;
```

```
if Education = "primary school" then d_Education=1;
else if Education = "secondary school" then d_Education=2;
else if Education = "college/bachelor degree" then d_Education=3;
else if Education = "masters degree" then d_Education=4;
run;
proc print;
run;

* logistic regression model;
proc logistic data=pred1;
model d_Finances(event='1')= Entertainment_spending Spending__on_looks d_Village_town;
output out=pred1 p=phat lower=lcl upper=ucl predprob1=(individual);
run;
proc print;
run;



* Find the final model, compute predicted value on training set, obtain the cut-off value for p;
proc logistic data=train1;
title "find cut-off";
model train_y1(event='1')= Entertainment_spending Spending__on_looks d_Village_town
/ctable pprob= (0.1 to 0.6 by 0.05);
*save predictions in sas dataset "pred";
output out=pred1(where=(train_y1=.))  p=phat lower=lcl upper=ucl
          predprob1=(individual);
run;

proc print data=pred1;
run;

/*Using cut-off value to compute classification matrix*/
data probs1;
set pred1;
pred_y1 = 0;
if (phat > 0.35) then pred_y1 = 1;
run;

proc print data=probs1;
run;

* compute classification matrix;
proc freq data=probs1;
tables d_Finances*pred_y1/norow nocol nopercent;
```

```
run;
```

Yesheng Qin section:

```
proc import datafile="S:\final\responses_QIN_YESHENG.csv" out=finance replace;
delimiter=',';
getnames=yes;
run;
proc print;
run;
*check distribution of Finances;
proc univariate data=finance;
var Finances;
histogram/normal (mu=est sigma=est);
run;

*delete rows with blank value to avoid inaccuracy;
data deal_blank;
set finance;
if cmiss(of _character_) + nmiss(of _numeric_) > 0 then delete;
run;

proc print data=deal_blank;
run;

*make the preditors into binary;
data binary_f;
set deal_blank;
Saver = 0;
if Finances = 4 then Saver=1;
else if Finances = 5 then Saver=1;
run;
```

```
*dummy variable for Gender;
data bin_gender;
set binary_f;
Gender1=1;
if Gender = 'female' then Gender1=0;
run;

 *dummy variable for Education;
 data bin_edu;
 set bin_gender;
 Edu_level=1;
 if Education = 'secondary school' then Edu_level=2;
 else if Education = 'college/bachelor degree' then Edu_level=3;
 else if Education = 'masters degree' then Edu_level=4;
 run;

*dummy variable for only_chilld;
data bin_child;
set bin_edu;
One_child=1;
if Only_child='no' then One_child=0;
run;


*dummy variable for VorC and HorB, this is the final dataset;
data data1;
set bin_child;
Village=1;
House=1;
if VorC='city' then Village=0;
if HorB='block of flats' then House=0;
run;


proc freq data= data1;
table Saver;
run;

*check interaction term;
proc logistic data=data1;
model Saver(evet='1') = Shp_centre Brd_cloth Enter_spd Spd_lk Spd_gadgt Spd_eat
Age Siblings Gender1 Edu_level One_child Village House Age*Edu_level Spd_lk*Brd_cloth/ stb
corrb;
```

```
run;

*fit full model with diagnostics to check conlinearity, outliers and influential points;
proc logistic data=data1;
model Saver(evet='1') = Shp_centre Brd_cloth Enter_spd Spd_lk Spd_gadgt Spd_eat
Age Siblings Gender1 Edu_level One_child Village House/ stb corrb influence iplots;
run;


*remove the first influential point;
data data2;
set data1;
if _n_=13 then delete;
run;

proc logistic data=data2;
model Saver(evet='1') = Shp_centre Brd_cloth Enter_spd Spd_lk Spd_gadgt Spd_eat
Age Siblings Gender1 Edu_level One_child Village House/ stb corrb influence iplots;
run;

*remove second influential point;
data data3;
set data2;
if _n_=722 then delete;
run;

proc logistic data=data3;
model Saver(evet='1') = Shp_centre Brd_cloth Enter_spd Spd_lk Spd_gadgt Spd_eat
Age Siblings Gender1 Edu_level One_child Village House/ stb corrb influence iplots;
run;

*check if the observation still enough;
proc freq data= data3;
table Saver;
run;

*prediction with full model;
data fullnew;
input Shp_centre Brd_cloth Enter_spd Spd_lk Spd_gadgt Spd_eat Age Siblings Gender1
Edu_level One_child Village House;
datalines;
1 2 2 2 1 2 27 2 1 4 0 0 1
4 5 3 5 2 4 22 0 0 3 1 1 0
```

```
;
data fullpred;
set fullnew data3;
run;
proc print data=fullpred;
run;
proc logistic data=fullpred;
model Saver(event='1')=Shp_centre Brd_cloth Enter_spd Spd_lk Spd_gadgt Spd_eat Age
Siblings Gender1 Edu_level One_child Village House;
output out=fullpred p=phat lower= lcl upper=ucl;
run;
proc print;
run;

*split into train and test set;
proc surveyselect data=data3 out=train seed=438821 samprate=0.65 outall;
run;
proc print;
run;

*compute new Y for train_y;
data train;
set train;
if Selected then train_y = Saver;
run;
proc print;
run;

*Model selection with training set (train_y);
*stepwise;
proc logistic data=train;
model train_y(evet='1') = Shp_centre Brd_cloth Enter_spd Spd_lk Spd_gadgt Spd_eat
Age Siblings Gender1 Edu_level One_child Village House Spd_lk*Brd_cloth/ selection=stepwise
rsquare;
run;

*backward;
proc logistic data=train;
model train_y(evet='1') = Shp_centre Brd_cloth Enter_spd Spd_lk Spd_gadgt Spd_eat
Age Siblings Gender1 Edu_level One_child Village House/ selection=backward rsquare;
run;

proc logistic data=data3;
```

```
model Saver(event='1')=Enter_spd Spd_lk/stb corrb influence iplots;
run;

data train_inter;
set train;
run;

*check if interaction is needed for the model;
proc logistic data=train_inter;
model train_y(event='1')=Enter_spd Spd_lk Enter_spd*Spd_lk/stb corrb influence iplots;
run;

*classification table and identify for the cutoff value;
proc logistic data=train;
model train_y(event='1') = Enter_spd Spd_lk/ ctable pprob=(0.1 to 0.65 by 0.05);
*0.6 has the highest value for pred.probability for test set;
output out=pred(where=(train_y=.)) p=phat lower=lcl uper=ucl;
run;



proc print data=pred;
run;

*cutoff value for classification matrics;
data probs;
set pred;
pred_y = 0;
*0.3 is from the pred.pro above;
if (phat > 0.30) then pred_y=1;
run;

proc print data=probs;
run;

*classification metrics;
proc freq data=probs;
table pred_y*Saver / norow nocol nopercent;
run;

*prediction with training set;
data new;
```

```
input Shp_centre Brd_cloth Enter_spd Spd_lk Spd_gadgt Spd_eat Age Siblings Gender $
Education $ Only_child $ VorC $ HorB $;
datalines;
. . 1 2 . . . . . . . . .
. . 5 5 . . . . . . . . .
;
data pred1;
set new train;
run;
proc print data=pred1;
run;

proc logistic data=pred1;
model train_y(event='1')=Enter_spd Spd_lk;
output out=pred1 p=phat lower= lcl upper=ucl;
run;
proc print;
run;
```

7. References

Manish Singh section:

Young People Survey, 2017. Explore the preferences, interests, habits, opinions, and fears of young people. www.kaggle.com. Accessed May 10, 2018.

https://stats.idre.ucla.edu/sas/dae/logit-regression/

Michal Chowaniak section:

Young People Survey, 2017. Explore the preferences, interests, habits, opinions, and fears of young people. www.kaggle.com. Accessed May 10, 2018.

Priyank Beno Cerejo section:

Young People Survey, 2017. Explore the preferences, interests, habits, opinions, and fears of young people. www.kaggle.com. Accessed May 10, 2018.

Rushabh Shah section:

Young People Survey, 2017. Explore the preferences, interests, habits, opinions, and fears of young people. www.kaggle.com. Accessed May 14, 2018.

Valentine Silvester Correia section:

Young People Survey, 2017. Explore the preferences, interests, habits, opinions, and fears of young people. www.kaggle.com. Accessed May 10, 2018.

Yesheng Qin Section:

Young People Survey, 2017. Explore the preferences, interests, habits, opinions, and fears of young people. www.kaggle.com. Accessed May 10, 2018.

Sullivan Bob, 2017. How the Recession Has Changed American Spending. blog.credit.com. Accessed June 1, 2018.