PORTFOLIO TASK 1 – LOGISTIC REGRESSION ANALYSIS

PART A

The aim of this report is to carry out analysis on behalf of a business analytics specialized consultancy on a subsample of weekly data from Fresco Supermarket, one of the biggest in UK. The team wants to identify trends and patterns in their sample of data which was collected over a 26-week period. The data includes customer's age, gender, shopping frequency per week and shopping basket value. The data also provides the information on consistency of customer's shopping basket regarding the type of products purchased. These can vary from value products, to brand as well as the supermarket's own high-quality product series Fresco Top. The data is collected from all the three types of stores collected by Fresco supermarket's team. They also want to identify if the spending potential of the customer falls in one of three categories namely: Low Spender (less than or equal to £25), Middle Spender (Between £25 and £70) and High Spender (Greater than £70).

As a business analyst, it is essential to understand the data; to determine if a model to predict any customer's shopping basket value is possible or not. Before the analysis, the data was rearranged, this will vary from software to software. The target variable is the spender category and it has three possible outcomes, thus, multinomial logistic regression is used to analyse the data. Using SPSS Statistics, MLR is performed to estimate a model to predict the spending potential of customer.

The reference category is 'Middle Spender'. The model is steered clear of assumptions first. Then a model is estimated to obtain parsimonious model after which it is tested for adequacy. Model's accuracy is found to be 82.7% which makes it a good model to be used for further analysis and predictions.

PART B

1. Variables

From the dataset given, i.e., Fresco.xls, it is clear that the variables to be analyzed are Customer ID, Shopping Basket, Gender, Age, Store Type, Value Products, Brand Products and

Top Fresco Products. The column 'Customer ID' will not be required as it is not a variable that will impact the analysis in any way. The remaining seven variables will be classified for further understanding.

Independent Variables:

The independent variables in this case are listed below.

- **Gender**: A categorical variable with two outcomes Male and Female.
- **Age**: A continuous variable showing the age of customer.
- **Store Type**: A categorical variable representing the three types of stores, namely, Superstores, Convenient stores and online stores.
- **Value Products**: A continuous variable showing the number of value products purchased by customers from Fresco.
- **Brand Products**: A continuous variable showing the number of brand products purchased by customers from Fresco.
- **Top Fresco Products**: A continuous variable showing the number of top Fresco products purchased by the customer from Fresco.

Dependent Variables:

In this case, the dependent variable is the 'Shopping Basket Value'.

- **Shopping Basket Value**: A categorical variable with three outcomes, i.e., Low Spender, Middle Spender and High Spender.

2. Multinomial Logistic Regression

From the given data, it is clear that the dependent variable is described by more than two possible outcomes. Therefore, the correct analysis method to be used will be the 'Multinomial Logistic Regression' using IBM SPSS Statistics. In this method of regression analysis, two binary logistics regression models are built and compared. If there are 'n' outcomes for the dependent variable, then there will be '(n-1)' binary logistic regression models; one for each of categories against the reference category. Thus, it is very important to identify the reference group to carry out the process. Generally, the one with the highest likelihood of occurrence is chosen as the reference group.

In this case, as there are three outcomes for the 'shopping basket value', there will be two binary logistic regression models. It can be found by running Descriptive statistics in SPSS statistics.

Shopping Basket

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Low Spender	18	24.0	24.0	24.0
	Middle Spender	30	40.0	40.0	64.0
	High Spender	27	36.0	36.0	100.0
	Total	75	100.0	100.0	

Fig 1: Frequency statistics

From the table above, as frequency of 'Middle Spender' is the highest, it will be the reference category.

3. Logistic Regression Assumptions

In this step Model's assumptions will be tested. If the assumptions are not satisfied, it will not be possible to build a parsimonious model.

1. Linearity

- To check for linearity, for each continuous variable a natural logarithm variable must be created
- After that, regression analysis should be performed with all the continuous variables with their natural logarithmic variables and interactions.
- Here, the significance of the coefficients of the interactions must be greater than 0.05.

Parameter Estimates

Shopping Basket[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
Low Spender	Intercept	-5702.113	3855.818	2.187	1	.139			
	Age	-651.489	403.646	2.605	1	.107	1.154E-283	.000	4.427E+60
	Value Products	11.999	77.974	.024	1	.878	162561.032	6.905E-62	3.827E+71
	Brand Products	386.270	286.617	1.816	1	.178	5.690E+167	6.116E-77	.[b]
	Top Fresco Products	-23.797	161.402	.022	1	.883	4.625E-11	1.902E-148	1.124E+127
	InAge	3856.908	2374.647	2.638	1	.104	.[b]	.000	.[b]
	InValueProducts	-65.758	186.140	.125	1	.724	2.764E-29	9.970E-188	7.665E+129
	InBrandProducts	-433.143	322.789	1.801	1	.180	7.734E-189	.000	4.437E+86
	InTopFrescoProducts	13.300	206.519	.004	1	.949	597165.477	9.692E-171	3.679E+181
	IntAge	116.871	72.832	2.575	1	.109	5.707E+50	5.781E-12	5.634E+112
	IntValueProducts	-1.142	17.792	.004	1	.949	.319	2.287E-16	4.454E+14
	IntBrandProducts	-110.841	82.791	1.792	1	.181	7.286E-49	2.457E-119	2.161E+22
	IntTopFrescoProducts	4.717	44.170	.011	1	.915	111.841	2.828E-36	4.423E+39
Middle Spender	Intercept	-1573.868	1564.424	1.012	1	.314			
	Age	-183.667	9.297	390.244	1	<.001	1.716E-80	2.091E-88	1.407E-72
	Value Products	78.156	70.547	1.227	1	.268	8.762E+33	7.809E-27	9.831E+93
	Brand Products	33.650	41.344	.662	1	.416	4.113E+14	2.644E-21	6.396E+49
	Top Fresco Products	-57.766	119.360	.234	1	.628	8.176E-26	2.057E-127	3.251E+76
	InAge	1146.807	498.250	5.298	1	.021	.[b]	8.713E+73	.[b]
	InValueProducts	-184.592	173.379	1.134	1	.287	6.804E-81	1.788E-228	2.589E+67
	InBrandProducts	-35.078	44.637	.618	1	.432	5.830E-16	5.901E-54	5.760E+22
	InTopFrescoProducts	46.040	177.597	.067	1	.795	9.880E+19	6.663E-132	1.465E+171
	IntAge	32.514	.000	.	1	.	1.321E+14	1.321E+14	1.321E+14
	IntValueProducts	-17.565	15.744	1.245	1	.265	2.353E-8	9.344E-22	592719.378
	IntBrandProducts	-8.942	11.083	.651	1	.420	.000	4.818E-14	355105.188
	IntTopFrescoProducts	15.717	30.474	.266	1	.606	6694199.025	7.697E-20	5.822E+32

a. The reference category is: High Spender.

b. Floating point overflow occurred while computing this statistic. Its value is therefore set to system missing.

Fig 2: Parameter Estimates with interactions

From the table, it is clear that the significance value of the interactions of all the continuous variables (IntAge, IntValueProducts, IntBrandProducts and IntTopFrescoProducts) are greater than 0.05.

Therefore, the test for linearity is satisfied.

2. <u>Independence of Errors</u>

The violation of this assumption can cause overdispersion. In such case, the independent variable can be falsely considered as significant.

- To test this assumption, it is necessary to ensure that the ratio between the chi-square goodness of fit and the degrees of freedom must be less than 2.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	16.632	112	1.000
Deviance	14.691	112	1.000

Fig 3: Goodness of fit

From the table, the required ratio can be calculated i.e., (16.632/112) =0.1485. The ratio value is less than 2, that means the assumption of independence of errors is also satisfied and the estimate model can be made.

4. Estimate Model

In this step, Multinomial Logistic Regression will be performed to build an initial model. Under 'Analyze' tab, the multinomial logistics option can be found in the 'Regression' option. The reference category is set to 'Middle Spender', as mentioned before. The dependent variable will be 'Shopping Basket value'. All the continuous variables will be under 'Covariates' (In SPSS metric independent variables are considered as covariates) and the categorical variables will be under 'Factors'.

Parameter Estimates

Shopping Basket[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B) Lower Bound	Upper Bound
Low Spender	Intercept	11.092	4.594	5.830	1	.016			
	Age	-.261	.169	2.382	1	.123	.770	.553	1.073
	Value Products	-.285	.208	1.878	1	.171	.752	.500	1.131
	Brand Products	-.420	.318	1.742	1	.187	.657	.352	1.226
	Top Fresco Products	-.351	.336	1.094	1	.296	.704	.364	1.359
High Spender	Intercept	-10.087	2.896	12.136	1	<.001			
	Age	.079	.047	2.865	1	.091	1.082	.988	1.186
	Value Products	.094	.078	1.427	1	.232	1.098	.942	1.281
	Brand Products	.126	.124	1.027	1	.311	1.134	.889	1.448
	Top Fresco Products	.428	.179	5.694	1	.017	1.534	1.079	2.181

a. The reference category is: Middle Spender.

Fig 4: Parameter Estimates – only variables

From the table above, the following information can be obtained.

- The reference category is set to Middle Spender.
- Top Fresco Products is the significant variable (as the values are 0.296 and 0.017). Even if one of the values is significant (value < 0.05), then that variable is considered significant.
- On comparing the other values, the insignificant variables are Age, Value Products and Brand Products.

The most parsimonious model should not contain insignificant variables. Therefore, to change this estimated model to a parsimonious model, the insignificant variables must be eliminated one by one until significant variables are obtained. It can be observed that the most insignificant variable is 'Brand Products' with a value of 0.187 in Low Spender category and 0.311 in High Spender category. Thus, the first variable to be eliminated is the Brand products.

5. Most parsimonious model

The same process is repeated as in the above section. The variable 'Brand Products' must be eliminated.

Parameter Estimates

Shopping Basket[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
Low Spender	Intercept	12.272	4.590	7.149	1	.008			
	Age	-.322	.155	4.333	1	.037	.724	.535	.981
	Value Products	-.352	.194	3.283	1	.070	.703	.481	1.029
	Top Fresco Products	-.582	.310	3.532	1	.060	.559	.305	1.025
High Spender	Intercept	-9.805	2.862	11.741	1	<.001			
	Age	.083	.046	3.258	1	.071	1.087	.993	1.190
	Value Products	.147	.068	4.653	1	.031	1.158	1.014	1.323
	Top Fresco Products	.421	.179	5.532	1	.019	1.524	1.073	2.165

a. The reference category is: Middle Spender.

Fig 5: Parameter estimates of final model

It can be observed from the table that all the variables are now significant. Hence, this model can be considered as a parsimonious model.

6. Tests for adequacy

1. Check whether the standardized residuals satisfy the conditions below.

- No more than 5% of absolute values are greater than 2.
- No more than 1% of absolute values are greater than 1.

From the image of .sav file below, it can be seen that the standardized residual values are less than 2, a couple values greater than 2 and the absolute values of remaining numbers are less than 1. Thus, the conditions for standardized residuals are satisfied.

	ZRE_1
1	2.1887
2	.1727
3	.
4	-.0869
5	.
6	-.2519
7	.
8	-.1270
9	.0058
10	.0000
11	.0000
12	.0196
13	.0025
14	-.4834
15	.0128
16	.0012
17	.0000
18	-.3519
19	.0079
20	.0214
21	2.9960
22	-.3212
23	.6463
24	.0117
25	-.2551
26	.
27	.0055
28	.
29	-.3928
30	-.7472
31	-.3702
32	.
33	.0000
34	.
35	.
36	.
37	.0000
38	.0115

	ZRE_2
1	-.0000
2	-.0622
3	.1426
4	.
5	.0117
6	.
7	.1051
8	.
9	-.3167
10	-.5260
11	-.6058
12	-.1232
13	-.3147
14	.
15	-.1754
16	-.5148
17	-1.2846
18	.
19	-.3773
20	-.1481
21	-.0000
22	.
23	-.0000
24	-.0000
25	.
26	1.6296
27	-.2216
28	.2280
29	.
30	.
31	.
32	.0319
33	-.5700
34	.1007
35	.0843
36	.1805
37	-.6246
38	-.2247
39	-.0000

Fig 6: ZRE_1 is standardized residuals values of Low Middle Spender. ZRE_2 is standardized residual values of Middle and High Spender.

2. Cook's distance should be less than 1.

	COO_1
1	.9683
2	.0076
3	.
4	.0004
5	.
6	.0148
7	.
8	.0007
9	.0000
10	.0000
11	.0000
12	.0000
13	.0000
14	.0851
15	.0000
16	.0000
17	.0000
18	.0428
19	.0000
20	.0000
21	1.0421
22	.0294
23	1.0314
24	.0000
25	.0421
26	.
27	.0000
28	.
29	.1019
30	.2174
31	.0301
32	.
33	.0000
34	.
35	.
36	.
37	.0000
38	.0000

	COO_2
1	.0000
2	.0000
3	.0018
4	.
5	.0000
6	.
7	.0007
8	.
9	.0183
10	.0868
11	.1119
12	.0006
13	.0159
14	.
15	.0016
16	.0691
17	1.0418
18	.
19	.0188
20	.0009
21	.0000
22	.
23	.0000
24	.0000
25	.
26	1.5910
27	.0030
28	.0051
29	.
30	.
31	.
32	.0000
33	.1111
34	.0006
35	.0001
36	.0023
37	.1850
38	.0049
39	.0000

Fig 7: COO_1 is Cook's distance for Low and Middle spender, COO_2 is Cook's distance for Middle and High Spender.

It can be seen that, except a couple values, all the values of Cook's distance are less than 1.

3. DFBeta values should be less than 1.

   The following image shows the DFBeta values of the case Low Spender and Middle Spender. It can be observed that almost all the values are less than 1.

DFB0_1	DFB1_1	DFB2_1	DFB3_1	DFB4_1	DFB5_1	DFB6_1	DFB7_1
-.86853	.53321	.04450	-.26572	.55775	.01366	-.07592	-.16649
-.37944	-.02140	.01652	-.52742	-.11245	.00817	.00896	.00539
.	.	.	.	.	.	.	.
-.04979	-.00046	.00041	-.09588	-.02662	.00613	.00108	.00547
.	.	.	.	.	.	.	.
-.59354	.07844	.01361	-.56826	-.02563	.02006	-.00374	.03948
.	.	.	.	.	.	.	.
-.07672	-.00470	.00152	-.12410	-.03412	.00522	.00437	.00571
-.00103	.00006	.00003	-.00145	-.00029	.00004	.00001	.00006
.00000	.00000	.00000	1.40683	.00000	.00000	.00000	.00000
.00000	.00000	.00000	.06642	.00000	.00000	.00000	.00000
-.00987	.00067	.00028	-.01476	-.00314	.00057	.00003	.00057
-.00024	.00001	.00001	-.00035	-.00007	.00001	.00000	.00002
-.02335	-.19088	-.00439	.77935	.43400	-.02113	.02072	-.03722
-.00438	.00000	.00014	-.00631	-.00126	.00021	-.00002	.00027
-.00007	.00000	.00000	-.00009	-.00002	.00000	.00000	.00000
.00000	.00000	.00000	.00115	.00000	.00000	.00000	.00000
-.71400	-.09910	.02364	-.42677	.11778	.00934	-.03079	.05133
-.00159	-.00001	.00005	-.00227	-.00045	.00006	.00001	.00012
-.01188	.00066	.00041	-.01598	-.00309	.00038	.00011	.00054
2.65208	.36918	-.07245	3.93229	1.24543	-.19513	-.07362	-.08184
-.79036	.14507	.02384	-.95767	-.22035	.01650	.00879	.04612
1.02279	.57320	-.11061	.30705	.22380	.01298	.50105	-.14920
-.00230	.00020	.00004	-.00376	-.00076	.00019	.00010	.00015

Fig 8: Image showing values of degrees of freedom – case 1

The following image shows the DFBeta values of the case Middle Spender and High Spender and all the values are less than 1.

DFB0_2	DFB1_2	DFB2_2	DFB3_2	DFB4_2	DFB5_2	DFB6_2	DFB7_2
.00000	.00000	.00000	.00000	-.00646	.00000	.00000	.00000
-.01853	.00126	.00014	.00050	.00492	.00010	.00018	.00089
-.06630	-.00503	.00080	.01777	.02954	.00283	-.00351	.06969
.	.	.	.	.	.	.	.
-.00079	-.00013	.00000	.00014	.00016	.00001	.00000	.00006
.	.	.	.	.	.	.	.
-.02315	-.00026	-.00013	.00656	.00202	.00172	-.00163	.00257
.	.	.	.	.	.	.	.
-.22449	-.12864	.00464	.09782	.18120	.00492	-.00946	.00235
-.45010	-.05541	.00584	-.21875	.14246	-.00446	.02095	.00737
-.16951	-.05289	-.00309	-.29554	-.06867	-.00060	.02130	.01876
-.05629	-.01352	.00072	.01283	.02209	-.00039	.00052	.00273
-.23142	-.09716	.00412	.09152	.13515	-.00248	.00032	.00751
.	.	.	.	.	.	.	.
-.09921	.01796	.00065	.00093	.02110	-.00090	.00335	.00440
-.32988	-.26285	.00810	.22059	.32006	.00117	-.00926	-.00059
-.62384	.38859	.01661	-.78535	.50149	.04174	-.08844	-.00896
.	.	.	.	.	.	.	.
-.31897	.02497	.00456	.02452	.18452	.00626	-.00316	-.00190
-.06390	-.02391	.00079	.01788	.03221	.00094	-.00112	.00281
.00000	.00000	.00000	.00000	-.01055	.00000	.00000	.00000
.	.	.	.	.	.	.	.
.00000	.00000	.00000	.00000	-.03134	.00000	.00000	.00000
.00000	.00000	.00000	.00000	-.53642	.00000	.00000	.00000
.	.	.	.	.	.	.	.

Fig 9: Image showing values of degrees of freedom – case 2

Therefore, the condition of DFBeta values is also satisfied.

4. Multicollinearity

- VIF should be less than 10.
- Tolerance statistic should be greater than 0.1.

From Linear regression, the test for multicollinearity can be done. On running the method, the following table will be obtained.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	.636	.161		3.945	<.001		
	Gender Value	.039	.101	.026	.390	.698	.781	1.280
	Age	.019	.005	.347	4.255	<.001	.510	1.959
	Store type value	.024	.064	.026	.376	.708	.688	1.452
	Value Products	.014	.006	.220	2.136	.036	.320	3.123
	Brand Products	.025	.013	.202	1.947	.056	.317	3.155
	Top Fresco Products	.037	.013	.253	2.743	.008	.398	2.510

a. Dependent Variable: Shopping Basket

Fig 10: Image showing statistics for test for multicollinearity

From the image above, we can see that all the values of the collinearity tolerance are greater than 0.1. The statistics VIF values are ranging from 1.280 to 3.155, they are less than 10. Therefore, it can be declared that the test for multicollinearity is also satisfied.

As the model has satisfied all the conditions, it can be said that this is an adequate model.

7. Goodness of fit

The Goodness of fit of a model can be tested using the Pseudo R square test, Hosmer and Lemeshow's test and Classification accuracy.

1. PSEUDO R SQUARE TEST

   For the model to pass this test, Cox and Snell and Nagelkerke's values must be close to 1.

Pseudo R-Square

Cox and Snell	.767
Nagelkerke	.868
McFadden	.676

Fig 11: Pseudo R Square

From the image above, it can be observed that both the values are close to 1. Hence, the model passes this test.

2. HOSMER AND LEMESHOW'S TEST

   For the model to pass this test, the significance value must be greater than 0.05 as the value compares the model predicted value to the real data.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	1.705	8	.989

Fig 12: Case 1: Low Spender and Middle Spender

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.845	8	.999

Fig 13: Case 2: Middle Spender and High Spender

In both the images, the Sig. value is greater than 0.05, therefore, the model passes this test.

3. CLASSIFICATION ACCURACY

In the classification table below, it can be seen that the overall percentage is 82.7%, which means that 82.7% cases were correctly classified. This number should be greater 70% for a good model.``

Classification

	Predicted			
Observed	Low Spender	Middle Spender	High Spender	Percent Correct
Low Spender	16	2	0	88.9%
Middle Spender	4	23	3	76.7%
High Spender	0	4	23	85.2%
Overall Percentage	26.7%	38.7%	34.7%	82.7%

Table Caption

Fig 14: Classification table of the dependent variable

Therefore, this model can be declared as the final model which will help identify in which category the next customer will fall.

The equations from which the probabilities can be calculated are as follows.

Model 1: Ln (ODDS) = 12.272 – 0.322*(Age) – 0.352*(Value Products) – 0.582*(Top Fresco Products)

e ^ Ln (ODDS) = [P (Low Spender)/P (Middle Spender)]

Model 2: Ln (ODDS) = -9.805 + 0.083*(Age) – 0.147*(Value Products) – 0.421*(Top Fresco Products)

e ^ Ln (ODDS) = [P (High Spender)/P (Middle Spender)]

P (Low Spender) + P (Middle Spender) + P (High Spender) = 1

8. Conclusion

Using Multinomial Logistic regression, analysis was carried out to predict the spending potential of a customer. Starting from choice of variables, assumptions were satisfied to be able to estimate a model and then parsimonious model was found and tested for adequacy. This led to the final model which can predict the category of the next customer.