Final Project - Logistic Regression: Shipping Data

Anjum Shams
DS423

Introduction

- Inspiration - personal experiences of packages being delayed.
- Goal and Technique - Develop a model to predict if the shipment will reach on time or not. Logistic Regression was used.
- Dataset - Shipping Dataset with 10999 observations.

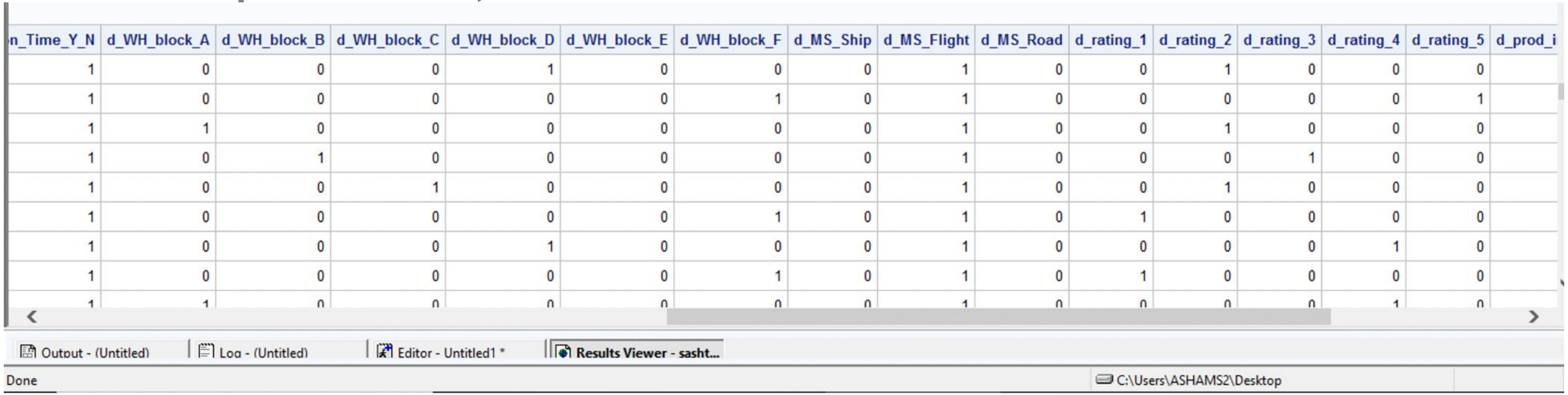
Dataset

- **Dependent Variable** - is Reached on time: where 1 Indicates that the product has NOT reached on time and 0 indicates it has reached on time.
- **11 Independent Variables** - ID, Warehouse_block, Mode_of _Shipment, Customer_care_calls, Customer_rating, Cost_of_the_Product, Prior_purchases, Product_importance, Gender, Discount_offered, Weight_in_gms.

New Dataset - Dummy Variables

- New Dataset - ID column dropped and
- Dummy variables were created for categorical variables:

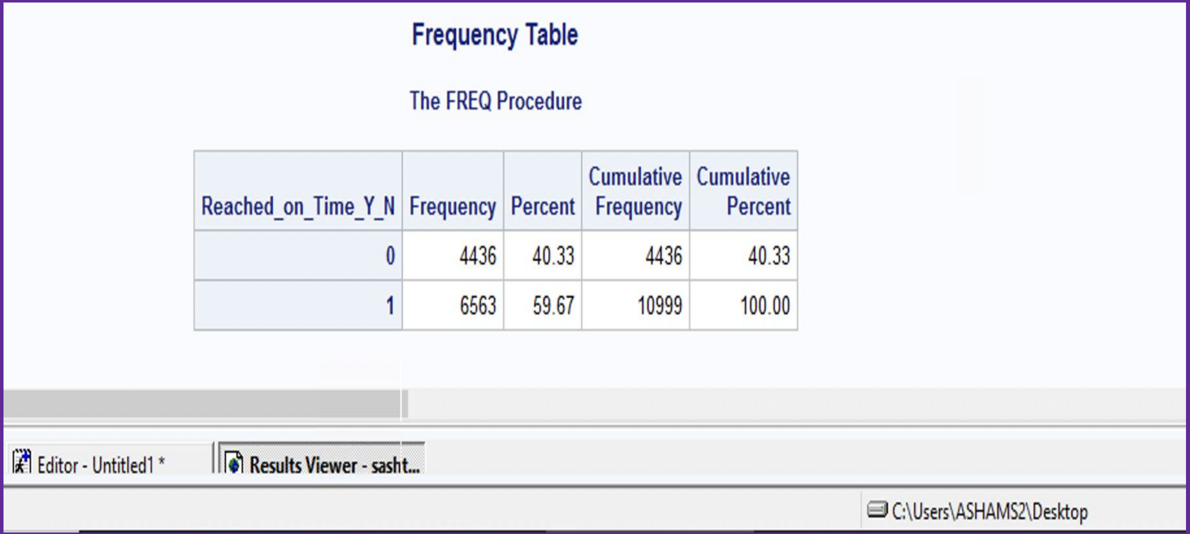
Warehouse_block, Mode_of _Shipment, Customer_rating, Product_importance, and Gender.



n_Time_Y_N	d_WH_block_A	d_WH_block_B	d_WH_block_C	d_WH_block_D	d_WH_block_E	d_WH_block_F	d_MS_Ship	d_MS_Flight	d_MS_Road	d_rating_1	d_rating_2	d_rating_3	d_rating_4	d_rating_5	d_prod_i
1	0	0	0	1	0	0	0	1	0	0	1	0	0	0	
1	0	0	0	0	0	1	0	1	0	0	0	0	0	1	
1	1	0	0	0	0	0	0	1	0	0	1	0	0	0	
1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	
1	0	0	1	0	0	0	0	1	0	0	1	0	0	0	
1	0	0	0	0	0	1	0	1	0	1	0	0	0	0	
1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	
1	0	0	0	0	0	1	0	1	0	1	0	0	0	0	
1	1	0	0	0	0	0	0	1	0	0	0	0	1	0	

Frequency Table

- The frequency table shows that the dataset is slightly imbalanced.



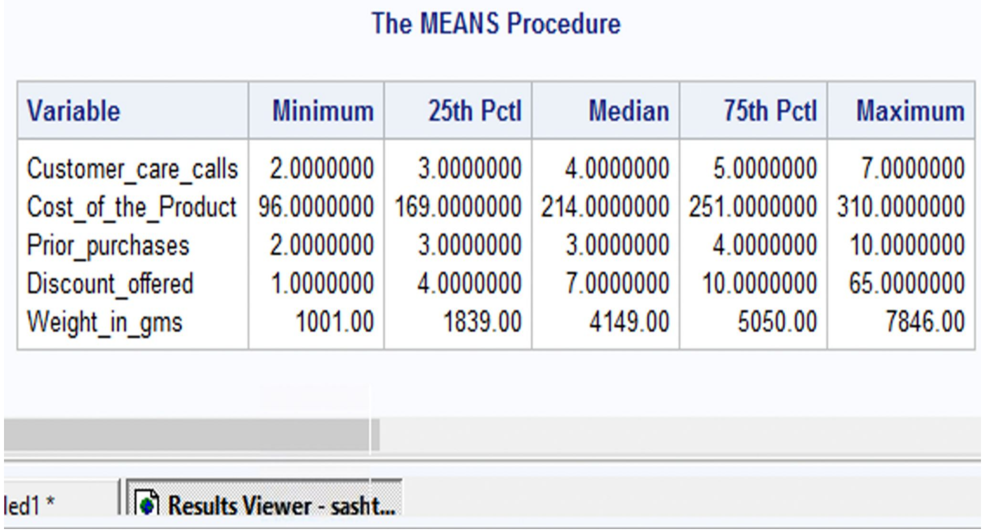
Frequency Table

The FREQ Procedure

Reached_on_Time_Y_N	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	4436	40.33	4436	40.33
1	6563	59.67	10999	100.00

Descriptives

- The median number of customer inquiry calls suggests delayed shipments and unsatisfied customers.
- The median discount suggests that on average less than 10 percent of discount was offered.

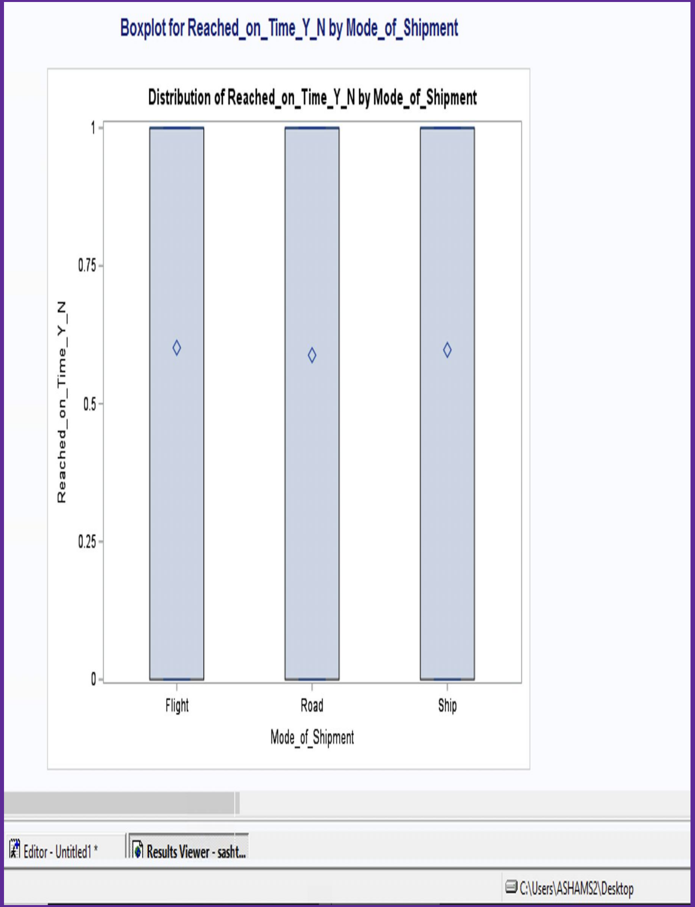


The MEANS Procedure

Variable	Minimum	25th Pctl	Median	75th Pctl	Maximum
Customer_care_calls	2.0000000	3.0000000	4.0000000	5.0000000	7.0000000
Cost_of_the_Product	96.0000000	169.0000000	214.0000000	251.0000000	310.0000000
Prior_purchases	2.0000000	3.0000000	3.0000000	4.0000000	10.0000000
Discount_offered	1.0000000	4.0000000	7.0000000	10.0000000	65.0000000
Weight_in_gms	1001.00	1839.00	4149.00	5050.00	7846.00

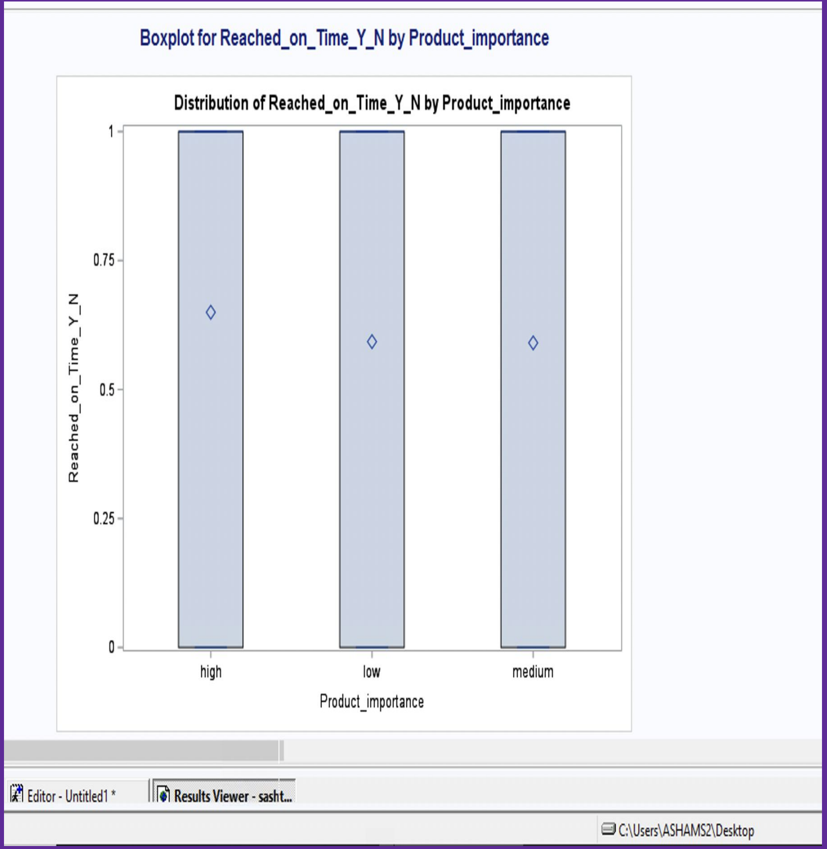
Box Plot

- The boxplot suggests that there is not much of a difference in the Flight, Road and Ship mode_of_shipment because the minimum, Q1, Q3, and max are almost the same.
- The Q2 is slightly higher for Flight, which could be due to the difference in the sample size.


Boxplot for Reached_on_Time_Y_N by Mode_of_Shipment
Distribution of Reached_on_Time_Y_N by Mode_of_Shipment
Reached_on_Time_Y_N
1
0.75
0.5
0.25
0
Flight
Road
Ship
Mode_of_Shipment
Editor - Untitled1 *
Results Viewer - sasht...
C:\Users\ASHAMS2\Desktop

Box Plot

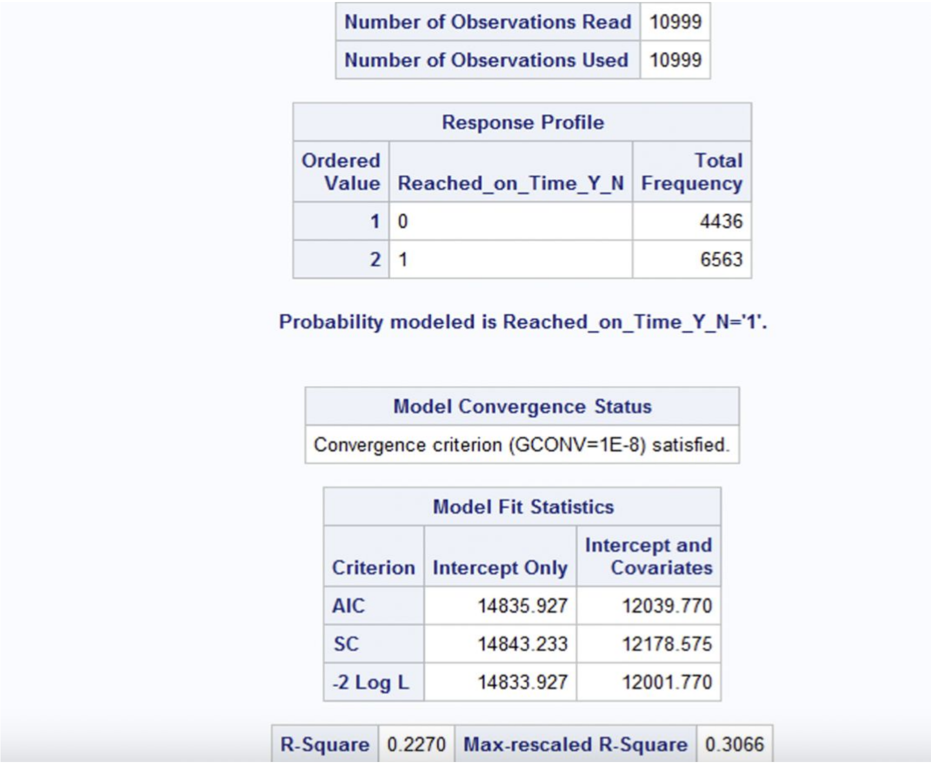
- The boxplot suggests that there is not much of a difference in the high, low, and medium Product_importance because the minimum, Q1, Q3, and maximum are almost the same.
- The Q2 is little higher for high Product_importance.


Boxplot for Reached_on_Time_Y_N by Product_importance
Distribution of Reached_on_Time_Y_N by Product_importance
Reached_on_Time_Y_N
1
0.75
0.5
0.25
0
high
low
medium
Product_importance
Editor - Untitled1 *
Results Viewer - sasht...
C:\Users\ASHAMS2\Desktop

Full Model Run with stb and Diagnostics

- The full model run shows R-Square value and AIC and SC values.

  The goal is to increase the R-Square value and decrease AIC and SC error term values.



Number of Observations Read	10999
Number of Observations Used	10999

Response Profile		
Ordered Value	Reached_on_Time_Y_N	Total Frequency
1	0	4436
2	1	6563

Probability modeled is Reached_on_Time_Y_N='1'.

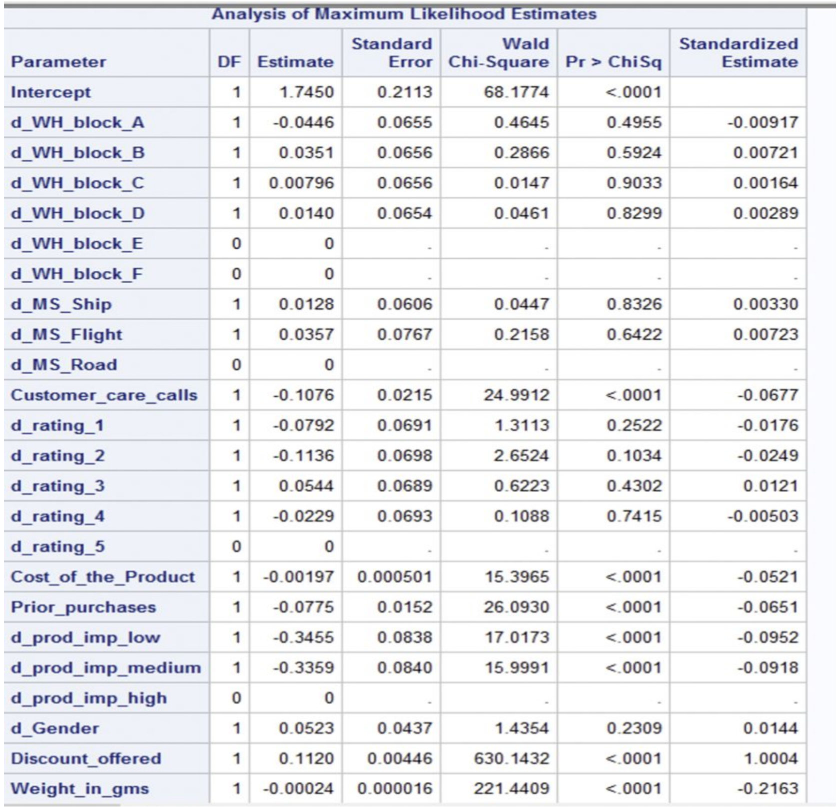
Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	14835.927	12039.770
SC	14843.233	12178.575
-2 Log L	14833.927	12001.770

R-Square	0.2270	Max-rescaled R-Square	0.3066

Standardized Estimates

Standardized Estimates of the independent Variables show that d_prod_imp_high, Discount_offered, Weight_in_gms, d_prod_imp_low, d_prod_imp_medium, Customer_care_calls, and Prior_purchases are most important predictors for the model.



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	1.7450	0.2113	68.1774	<.0001	
d_WH_block_A	1	-0.0446	0.0655	0.4645	0.4955	-0.00917
d_WH_block_B	1	0.0351	0.0656	0.2866	0.5924	0.00721
d_WH_block_C	1	0.00796	0.0656	0.0147	0.9033	0.00164
d_WH_block_D	1	0.0140	0.0654	0.0461	0.8299	0.00289
d_WH_block_E	0	0	.	.	.	.
d_WH_block_F	0	0	.	.	.	.
d_MS_Ship	1	0.0128	0.0606	0.0447	0.8326	0.00330
d_MS_Flight	1	0.0357	0.0767	0.2158	0.6422	0.00723
d_MS_Road	0	0	.	.	.	.
Customer_care_calls	1	-0.1076	0.0215	24.9912	<.0001	-0.0677
d_rating_1	1	-0.0792	0.0691	1.3113	0.2522	-0.0176
d_rating_2	1	-0.1136	0.0698	2.6524	0.1034	-0.0249
d_rating_3	1	0.0544	0.0689	0.6223	0.4302	0.0121
d_rating_4	1	-0.0229	0.0693	0.1088	0.7415	-0.00503
d_rating_5	0	0	.	.	.	.
Cost_of_the_Product	1	-0.00197	0.000501	15.3965	<.0001	-0.0521
Prior_purchases	1	-0.0775	0.0152	26.0930	<.0001	-0.0651
d_prod_imp_low	1	-0.3455	0.0838	17.0173	<.0001	-0.0952
d_prod_imp_medium	1	-0.3359	0.0840	15.9991	<.0001	-0.0918
d_prod_imp_high	0	0	.	.	.	.
d_Gender	1	0.0523	0.0437	1.4354	0.2309	0.0144
Discount_offered	1	0.1120	0.00446	630.1432	<.0001	1.0004
Weight_in_gms	1	-0.00024	0.000016	221.4409	<.0001	-0.2163

Full Model Equation

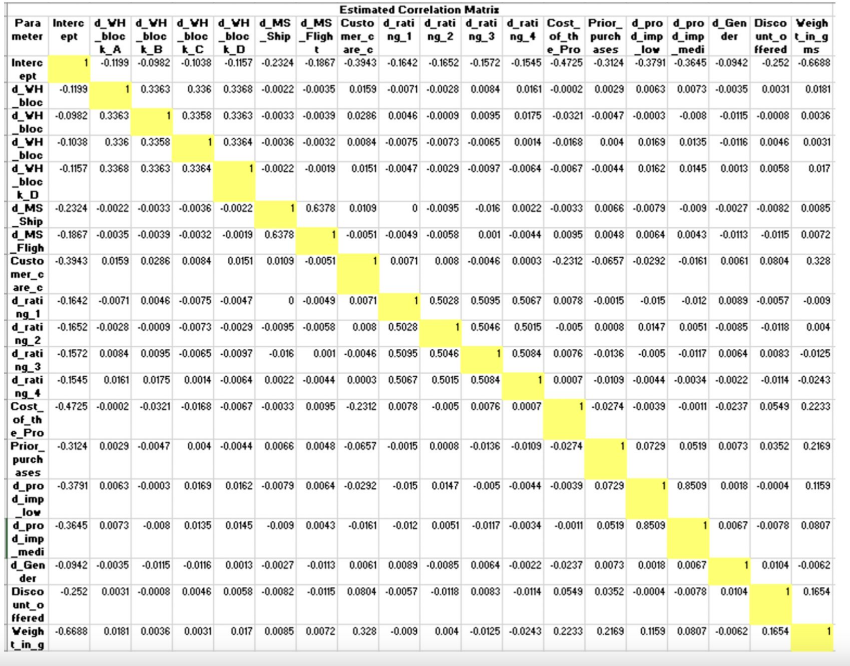
Full Model equation:

The full logistic regression model to predict probability of reached on time p=Pr(reached_on_time=1) is fitted using PROC LOGISTIC:

log(p/1-p) = 1.75 - 0.04 d_WH_block_A +0.04 d_WH_block_B + 0.008 d_WH_block_C + 0.01 d_WH_block_D + 0.01d_MS_Ship + 0.03 d_MS_Flight -0.1 Customer_care_calls - 0.08 d_rating_1 - 0.11 d_rating_2 + 0.05 d_rating_3-0.02 d_rating_4 - 0.002 Cost_of_the_Product -0.08 Prior_purchases - 0.3 d_prod_imp_low - 0.3 d_prod_imp_medium +0.05 d_Gender + 0.11 Discount_offered - 0.0002 Weight_in_gms

Full model - Multicollinearity Scan

- The values above and below diagonal in the estimated correlation matrix are < 0.9.
- Therefore, there is no multicollinearity among the independent variables.

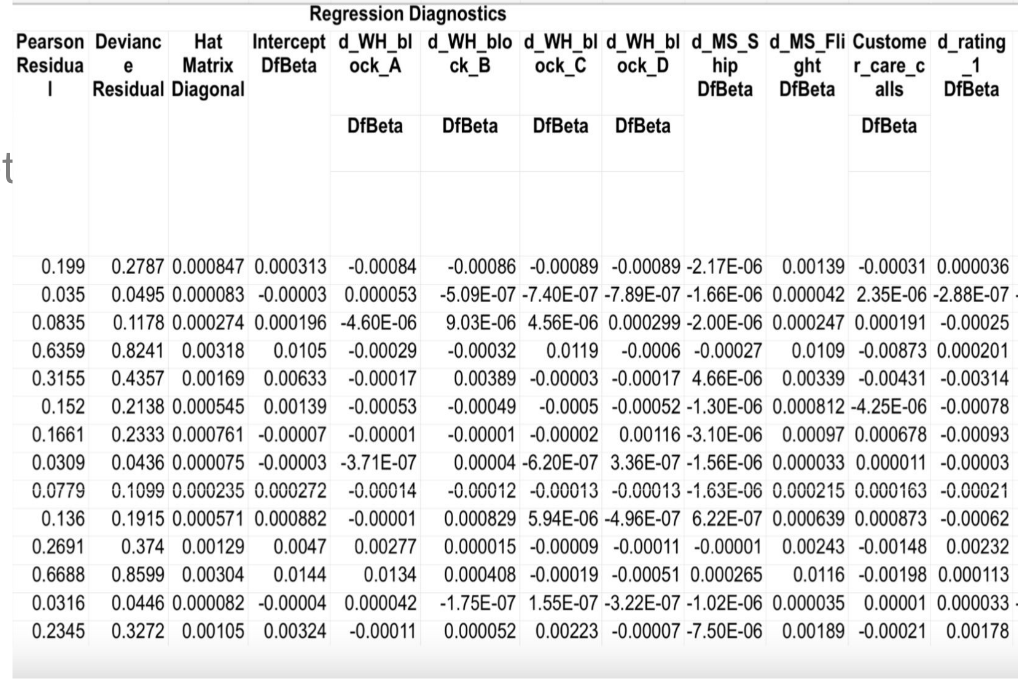


Estimated Correlation Matrix

Parameter	Intercept	d_VH_block_A	d_VH_block_B	d_VH_block_C	d_VH_block_D	d_MS_Ship	d_MS_Flight	Customer_care_c	d_rating_1	d_rating_2	d_rating_3	d_rating_4	Cost_of_the_Pro	Prior_purchases	d_prod_imp_low	d_prod_imp_medi	d_Gender	Discount_offered	Weight_in_gms
Intercept	1	-0.1199	-0.0982	-0.1038	-0.1157	-0.2324	-0.1867	-0.3943	-0.1642	-0.1652	-0.1572	-0.1545	-0.4725	-0.3124	-0.3791	-0.3645	-0.0942	-0.252	-0.6688
d_VH_bloc	-0.1199	1	0.3363	0.336	0.3368	-0.0022	-0.0035	0.0159	-0.0071	-0.0028	0.0084	0.0161	-0.0002	0.0029	0.0063	0.0073	-0.0035	0.0031	0.0181
d_VH_bloc	-0.0982	0.3363	1	0.3358	0.3363	-0.0033	-0.0039	0.0286	0.0046	-0.0009	0.0095	0.0175	-0.0321	-0.0047	-0.0003	-0.008	-0.0115	-0.0008	0.0036
d_VH_bloc	-0.1038	0.336	0.3358	1	0.3364	-0.0036	-0.0032	0.0084	-0.0075	-0.0073	-0.0065	0.0014	-0.0168	0.004	0.0169	0.0135	-0.0116	0.0046	0.0031
d_VH_block_D	-0.1157	0.3368	0.3363	0.3364	1	-0.0022	-0.0019	0.0151	-0.0047	-0.0029	-0.0097	-0.0064	-0.0067	-0.0044	0.0162	0.0145	0.0013	0.0058	0.017
d_MS_Ship	-0.2324	-0.0022	-0.0033	-0.0036	-0.0022	1	0.6378	0.0109	0	-0.0095	-0.016	0.0022	-0.0033	0.0066	-0.0079	-0.009	-0.0027	-0.0082	0.0085
d_MS_Fligh	-0.1867	-0.0035	-0.0039	-0.0032	-0.0019	0.6378	1	-0.0051	-0.0049	-0.0058	0.001	-0.0044	0.0095	0.0048	0.0064	0.0043	-0.0113	-0.0115	0.0072
Customer_care_c	-0.3943	0.0159	0.0286	0.0084	0.0151	0.0109	-0.0051	1	0.0071	0.008	-0.0046	0.0003	-0.2312	-0.0657	-0.0292	-0.0161	0.0061	0.0804	0.328
d_rating_1	-0.1642	-0.0071	0.0046	-0.0075	-0.0047	0	-0.0049	0.0071	1	0.5028	0.5095	0.5067	0.0078	-0.0015	-0.015	-0.012	0.0089	-0.0057	-0.009
d_rating_2	-0.1652	-0.0028	-0.0009	-0.0073	-0.0029	-0.0095	-0.0058	0.008	0.5028	1	0.5046	0.5015	-0.005	0.0008	0.0147	0.0051	-0.0085	-0.0118	0.004
d_rating_3	-0.1572	0.0084	0.0095	-0.0065	-0.0097	-0.016	0.001	-0.0046	0.5095	0.5046	1	0.5084	0.0076	-0.0136	-0.005	-0.0117	0.0064	0.0083	-0.0125
d_rating_4	-0.1545	0.0161	0.0175	0.0014	-0.0064	0.0022	-0.0044	0.0003	0.5067	0.5015	0.5084	1	0.0007	-0.0109	-0.0044	-0.0034	-0.0022	-0.0114	-0.0243
Cost_of_the_Pro	-0.4725	-0.0002	-0.0321	-0.0168	-0.0067	-0.0033	0.0095	-0.2312	0.0078	-0.005	0.0076	0.0007	1	-0.0274	-0.0039	-0.0011	-0.0237	0.0549	0.2233
Prior_purchases	-0.3124	0.0029	-0.0047	0.004	-0.0044	0.0066	0.0048	-0.0657	-0.0015	0.0008	-0.0136	-0.0109	-0.0274	1	0.0729	0.0519	0.0073	0.0352	0.2169
d_prod_imp_low	-0.3791	0.0063	-0.0003	0.0169	0.0162	-0.0079	0.0064	-0.0292	-0.015	0.0147	-0.005	-0.0044	-0.0039	0.0729	1	0.8509	0.0018	-0.0004	0.1159
d_prod_imp_medi	-0.3645	0.0073	-0.008	0.0135	0.0145	-0.009	0.0043	-0.0161	-0.012	0.0051	-0.0117	-0.0034	-0.0011	0.0519	0.8509	1	0.0067	-0.0078	0.0807
d_Gender	-0.0942	-0.0035	-0.0115	-0.0116	0.0013	-0.0027	-0.0113	0.0061	0.0089	-0.0085	0.0064	-0.0022	-0.0237	0.0073	0.0018	0.0067	1	0.0104	-0.0062
Discount_offered	-0.252	0.0031	-0.0008	0.0046	0.0058	-0.0082	-0.0115	0.0804	-0.0057	-0.0118	0.0083	-0.0114	0.0549	0.0352	-0.0004	-0.0078	0.0104	1	0.1654
Weight_in_g	-0.6688	0.0181	0.0036	0.0031	0.017	0.0085	0.0072	0.328	-0.009	0.004	-0.0125	-0.0243	0.2233	0.2169	0.1159	0.0807	-0.0062	0.1654	1

Full model - Outliers and Influential points Scan

- The deviance residual plot shows no outliers (>=+3 or -3).
- For influential points scan for |Dfbet 2/sqrt(n)= 2/sqrt(10999)=0.019
- After scanning for |Dfbeta| > 0.019, There were multiple influential points detected and some were removed to see the improvement in the model.



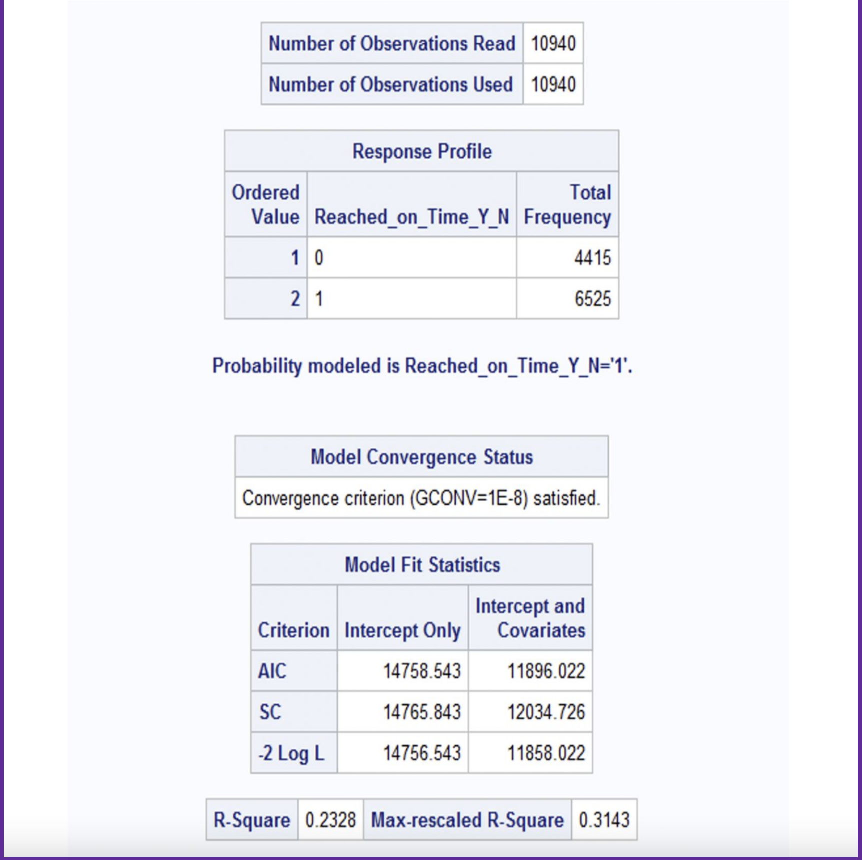
Regression Diagnostics

Pearson Residual	Deviance Residual	Hat Matrix Diagonal	Intercept DfBeta	d_WH_block_A DfBeta	d_WH_block_B DfBeta	d_WH_block_C DfBeta	d_WH_block_D DfBeta	d_MS_Ship DfBeta	d_MS_Flight DfBeta	Customer_care_calls DfBeta	d_rating_1 DfBeta
0.199	0.2787	0.000847	0.000313	-0.00084	-0.00086	-0.00089	-0.00089	-2.17E-06	0.00139	-0.00031	0.000036
0.035	0.0495	0.000083	-0.00003	0.000053	-5.09E-07	-7.40E-07	-7.89E-07	-1.66E-06	0.000042	2.35E-06	-2.88E-07
0.0835	0.1178	0.000274	0.000196	-4.60E-06	9.03E-06	4.56E-06	0.000299	-2.00E-06	0.000247	0.000191	-0.00025
0.6359	0.8241	0.00318	0.0105	-0.00029	-0.00032	0.0119	-0.0006	-0.00027	0.0109	-0.00873	0.000201
0.3155	0.4357	0.00169	0.00633	-0.00017	0.00389	-0.00003	-0.00017	4.66E-06	0.00339	-0.00431	-0.00314
0.152	0.2138	0.000545	0.00139	-0.00053	-0.00049	-0.0005	-0.00052	-1.30E-06	0.000812	-4.25E-06	-0.00078
0.1661	0.2333	0.000761	-0.00007	-0.00001	-0.00001	-0.00002	0.00116	-3.10E-06	0.00097	0.000678	-0.00093
0.0309	0.0436	0.000075	-0.00003	-3.71E-07	0.00004	-6.20E-07	3.36E-07	-1.56E-06	0.000033	0.000011	-0.00003
0.0779	0.1099	0.000235	0.000272	-0.00014	-0.00012	-0.00013	-0.00013	-1.63E-06	0.000215	0.000163	-0.00021
0.136	0.1915	0.000571	0.000882	-0.00001	0.000829	5.94E-06	-4.96E-07	6.22E-07	0.000639	0.000873	-0.00062
0.2691	0.374	0.00129	0.0047	0.00277	0.000015	-0.00009	-0.00011	-0.00001	0.00243	-0.00148	0.00232
0.6688	0.8599	0.00304	0.0144	0.0134	0.000408	-0.00019	-0.00051	0.000265	0.0116	-0.00198	0.000113
0.0316	0.0446	0.000082	-0.00004	0.000042	-1.75E-07	1.55E-07	-3.22E-07	-1.02E-06	0.000035	0.00001	0.000033
0.2345	0.3272	0.00105	0.00324	-0.00011	0.000052	0.00223	-0.00007	-7.50E-06	0.00189	-0.00021	0.00178

Full Model - After Removing Influential Points

- There is not much improvement in the model after removing the Influential points.

- The R-Square value has improved only by 0.01.



Number of Observations Read	10940
Number of Observations Used	10940

Response Profile

Ordered Value	Reached_on_Time_Y_N	Total Frequency
1	0	4415
2	1	6525

Probability modeled is Reached_on_Time_Y_N='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

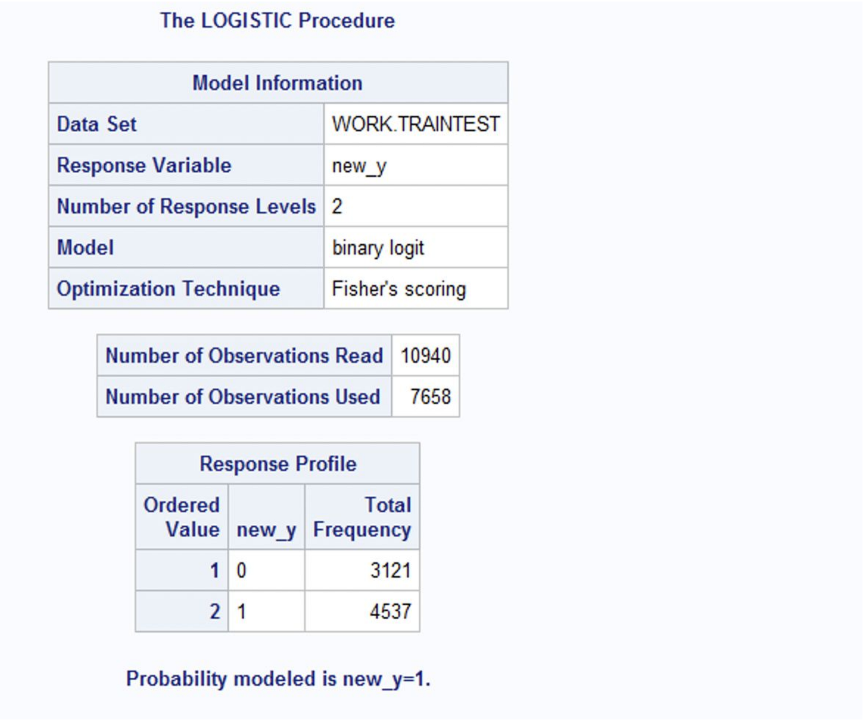
Criterion	Intercept Only	Intercept and Covariates
AIC	14758.543	11896.022
SC	14765.843	12034.726
-2 Log L	14756.543	11858.022

R-Square	0.2328	Max-rescaled R-Square	0.3143

Data Split into Train and Test Sets

- The data was split into 70:30 ratio.

- 70% of the data was used to train the model and 30% was used to test the model.



The LOGISTIC Procedure

Model Information	
Data Set	WORK.TRAINTEST
Response Variable	new_y
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	10940
Number of Observations Used	7658

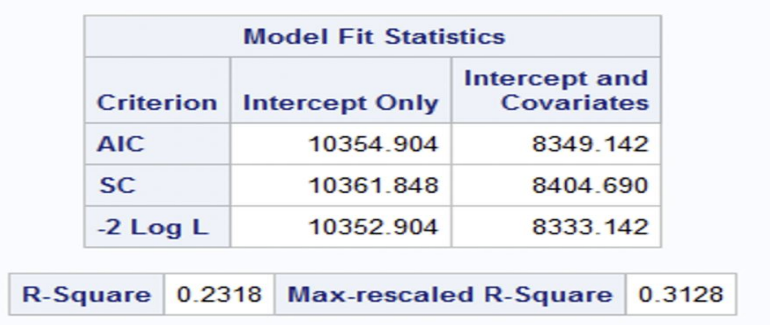
Response Profile

Ordered Value	new_y	Total Frequency
1	0	3121
2	1	4537

Probability modeled is new_y=1.

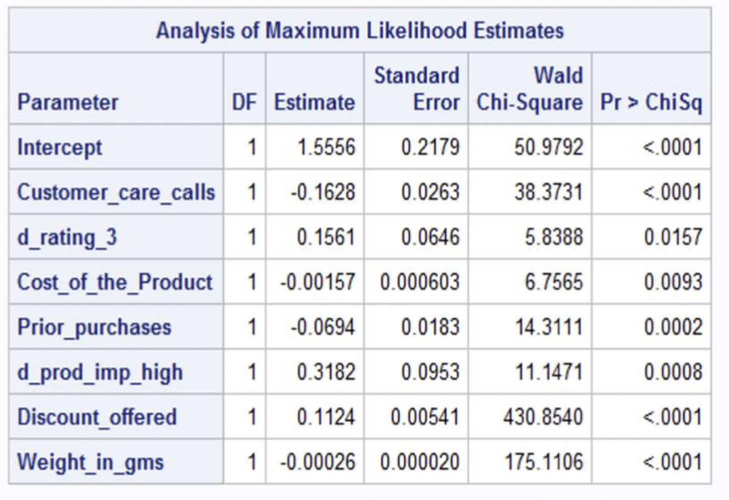
Stepwise Selection Method

- The full regression model using stepwise selection method resulted in a model with 7 significant predictors, and a R-Square value of 0.23.



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	10354.904	8349.142
SC	10361.848	8404.690
-2 Log L	10352.904	8333.142

R-Square	0.2318	Max-rescaled R-Square	0.3128



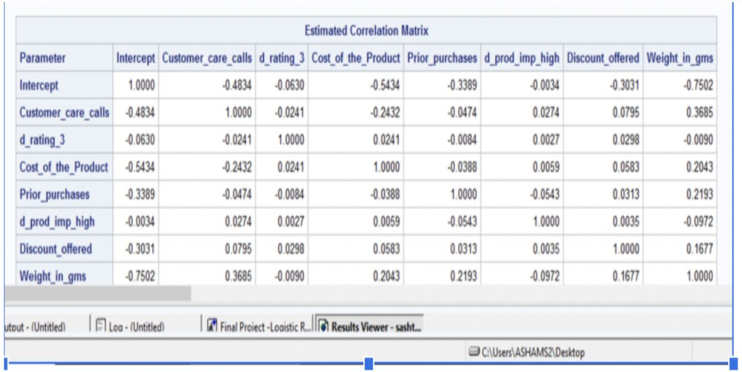
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.5556	0.2179	50.9792	<.0001
Customer_care_calls	1	-0.1628	0.0263	38.3731	<.0001
d_rating_3	1	0.1561	0.0646	5.8388	0.0157
Cost_of_the_Product	1	-0.00157	0.000603	6.7565	0.0093
Prior_purchases	1	-0.0694	0.0183	14.3111	0.0002
d_prod_imp_high	1	0.3182	0.0953	11.1471	0.0008
Discount_offered	1	0.1124	0.00541	430.8540	<.0001
Weight_in_gms	1	-0.00026	0.000020	175.1106	<.0001

Final Model with Diagnostics

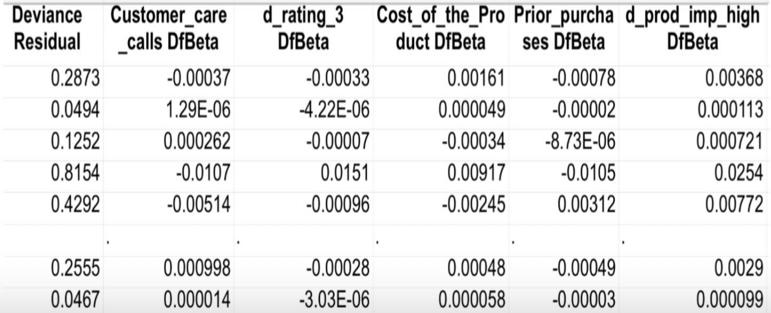
- The final model was run with diagnostics, there were no issues of multicollinearity detected. The deviance residuals showed no outliers.
- There were several influential points detected, some were removed, but not much Improvement was shown in the model after removing the Influential points.

Multicollinearity scan



Estimated Correlation Matrix								
Parameter	Intercept	Customer_care_calls	d_rating_3	Cost_of_the_Product	Prior_purchases	d_prod_imp_high	Discount_offered	Weight_in_gms
Intercept	1.0000	-0.4834	-0.0630	-0.5434	-0.3389	-0.0034	-0.3031	-0.7502
Customer_care_calls	-0.4834	1.0000	-0.0241	-0.2432	-0.0474	0.0274	0.0795	0.3685
d_rating_3	-0.0630	-0.0241	1.0000	0.0241	-0.0084	0.0027	0.0298	-0.0090
Cost_of_the_Product	-0.5434	-0.2432	0.0241	1.0000	-0.0388	0.0059	0.0583	0.2043
Prior_purchases	-0.3389	-0.0474	-0.0084	-0.0388	1.0000	-0.0543	0.0313	0.2193
d_prod_imp_high	-0.0034	0.0274	0.0027	0.0059	-0.0543	1.0000	0.0035	-0.0972
Discount_offered	-0.3031	0.0795	0.0298	0.0583	0.0313	0.0035	1.0000	0.1677
Weight_in_gms	-0.7502	0.3685	-0.0090	0.2043	0.2193	-0.0972	0.1677	1.0000

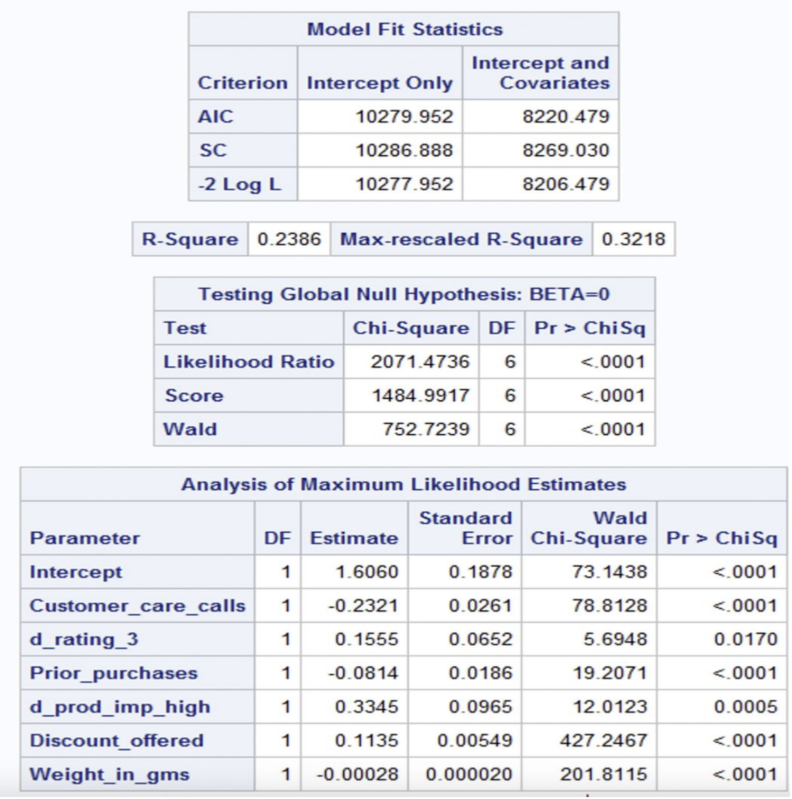
Outliers and influential points scan



Deviance Residual	Customer_care _calls DfBeta	d_rating_3 DfBeta	Cost_of_the_Pro duct DfBeta	Prior_purcha ses DfBeta	d_prod_imp_high DfBeta
0.2873	-0.00037	-0.00033	0.00161	-0.00078	0.00368
0.0494	1.29E-06	-4.22E-06	0.000049	-0.00002	0.000113
0.1252	0.000262	-0.00007	-0.00034	-8.73E-06	0.000721
0.8154	-0.0107	0.0151	0.00917	-0.0105	0.0254
0.4292	-0.00514	-0.00096	-0.00245	0.00312	0.00772
.	.	.	.	.	
0.2555	0.000998	-0.00028	0.00048	-0.00049	0.0029
0.0467	0.000014	-3.03E-06	0.000058	-0.00003	0.000099

Final model

- In the final model, one of the predictors was found insignificant and was removed.
- The final model has an improved R-square value of 0.2386 and 6 significant predictors.
- 23.86% of the variation in reached_on_time is explained by the model, the rest is unexplained.
- **Final model equation:**
  log(reached_on_time =1/reached_on_time=0) = 1.61
  - 0.23 Customer_care_calls + 0.16 d_rating_3
  - 0.08 Prior_purchases + 0.33 d_prod_imp_high + 0.11
  Discount_offered - 0.00028 Weight_in_gms
- Customer_care_calls = [exp(-0.23)-1]*100 = -20.55%
  If Customer_care_calls increases by 1 call reached_on_t
  will decrease by 20.55%



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	10279.952	8220.479
SC	10286.888	8269.030
-2 Log L	10277.952	8206.479

R-Square	0.2386	Max-rescaled R-Square	0.3218

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2071.4736	6	<.0001
Score	1484.9917	6	<.0001
Wald	752.7239	6	<.0001

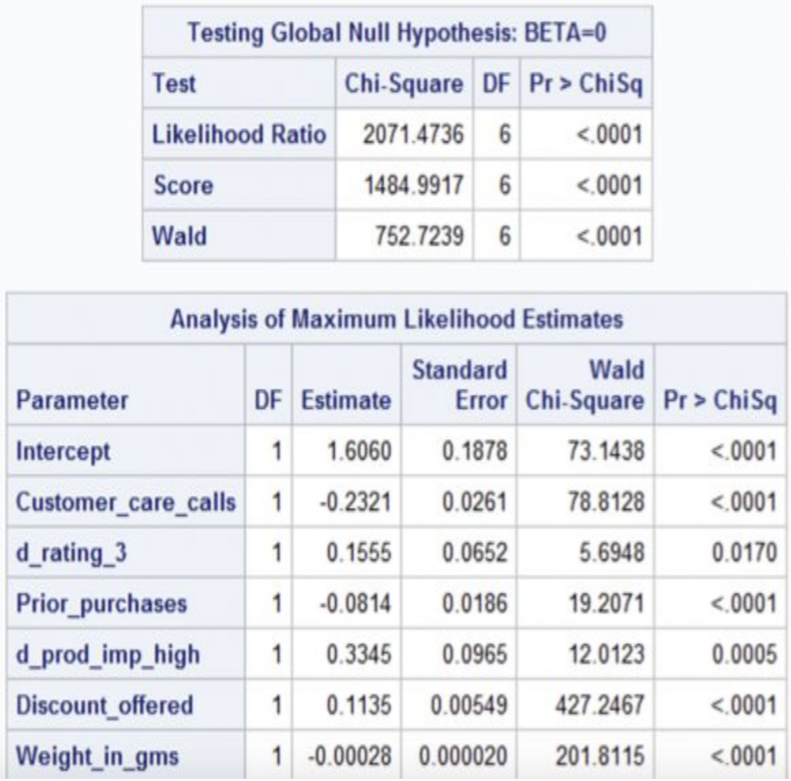
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.6060	0.1878	73.1438	<.0001
Customer_care_calls	1	-0.2321	0.0261	78.8128	<.0001
d_rating_3	1	0.1555	0.0652	5.6948	0.0170
Prior_purchases	1	-0.0814	0.0186	19.2071	<.0001
d_prod_imp_high	1	0.3345	0.0965	12.0123	0.0005
Discount_offered	1	0.1135	0.00549	427.2467	<.0001
Weight_in_gms	1	-0.00028	0.000020	201.8115	<.0001

Final model - Goodness of Fit Test

Goodness of Fit test:

- Ho:βj=0 Ha:βj≠0
- Likelihood Ratio = 2071.4736 P-value = <0.0001
- Conclusion: We can reject the null hypothesis because

P-value is very small, less than alpha = 0.05. This means there is at least one significant predictor which has a strong association with Y. The F-test gives a strong support to the fitted model.



Testing Global Null Hypothesis: BETA=0

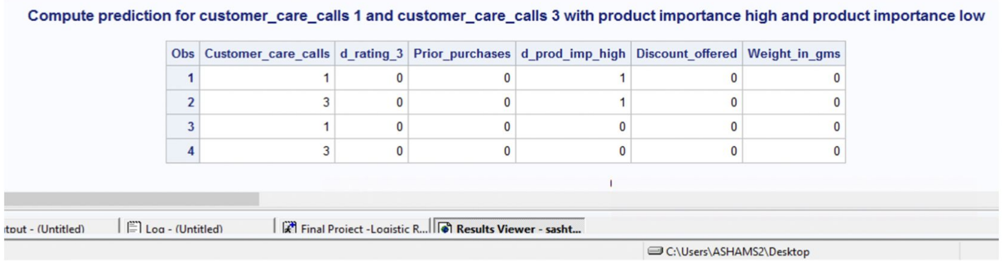
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2071.4736	6	<.0001
Score	1484.9917	6	<.0001
Wald	752.7239	6	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.6060	0.1878	73.1438	<.0001
Customer_care_calls	1	-0.2321	0.0261	78.8128	<.0001
d_rating_3	1	0.1555	0.0652	5.6948	0.0170
Prior_purchases	1	-0.0814	0.0186	19.2071	<.0001
d_prod_imp_high	1	0.3345	0.0965	12.0123	0.0005
Discount_offered	1	0.1135	0.00549	427.2467	<.0001
Weight_in_gms	1	-0.00028	0.000020	201.8115	<.0001

Compute Predictions and Merge Data

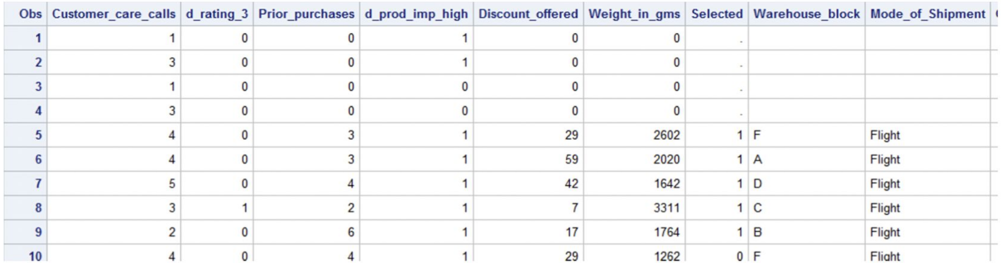
- Predictions were computed for Customer_care_calls 1 and Customer_care_calls 3 with Product_importance_high And Product_importance_low.
- The new dataset was merged with the original dataset.



Compute prediction for customer_care_calls 1 and customer_care_calls 3 with product importance high and product importance low

Obs	Customer_care_calls	d_rating_3	Prior_purchases	d_prod_imp_high	Discount_offered	Weight_in_gms
1	1	0	0	1	0	0
2	3	0	0	1	0	0
3	1	0	0	0	0	0
4	3	0	0	0	0	0

Merged data



Obs	Customer_care_calls	d_rating_3	Prior_purchases	d_prod_imp_high	Discount_offered	Weight_in_gms	Selected	Warehouse_block	Mode_of_Shipment
1	1	0	0	1	0	0	.		
2	3	0	0	1	0	0	.		
3	1	0	0	0	0	0	.		
4	3	0	0	0	0	0	.		
5	4	0	3	1	29	2602	1	F	Flight
6	4	0	3	1	59	2020	1	A	Flight
7	5	0	4	1	42	1642	1	D	Flight
8	3	1	2	1	7	3311	1	C	Flight
9	2	0	6	1	17	1764	1	B	Flight
10	4	0	4	1	29	1262	0	F	Flight

Predictive Probabilities and Confidence Intervals

- Phat = 0.847 = 84.7%
- Lcl = [exp(0.79)-1]*100 = 120.34%
- Ucl = [exp(0.88)-1]*100 =141.08%
- If a shipment has Customer_care_calls

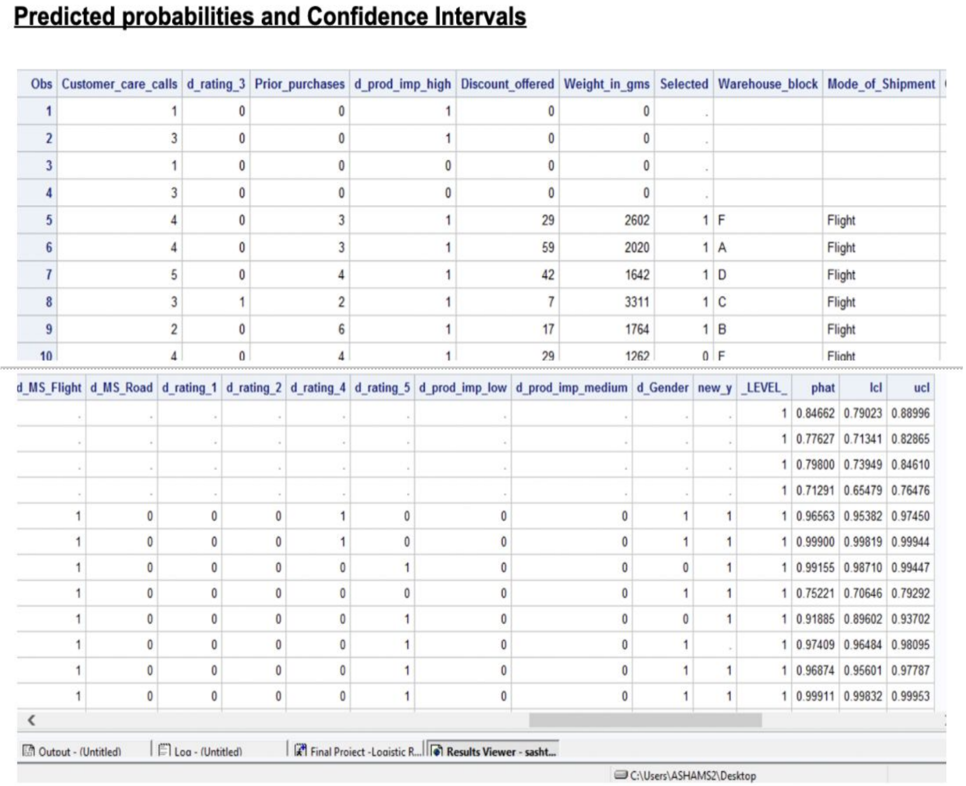
  = 1, and d_prod_imp_high = 1.

  The predicted probability of

  Reached_on_time is 84.7%.

- it is expected to fall within the range of

  120.34% - 141.08% confidence interval



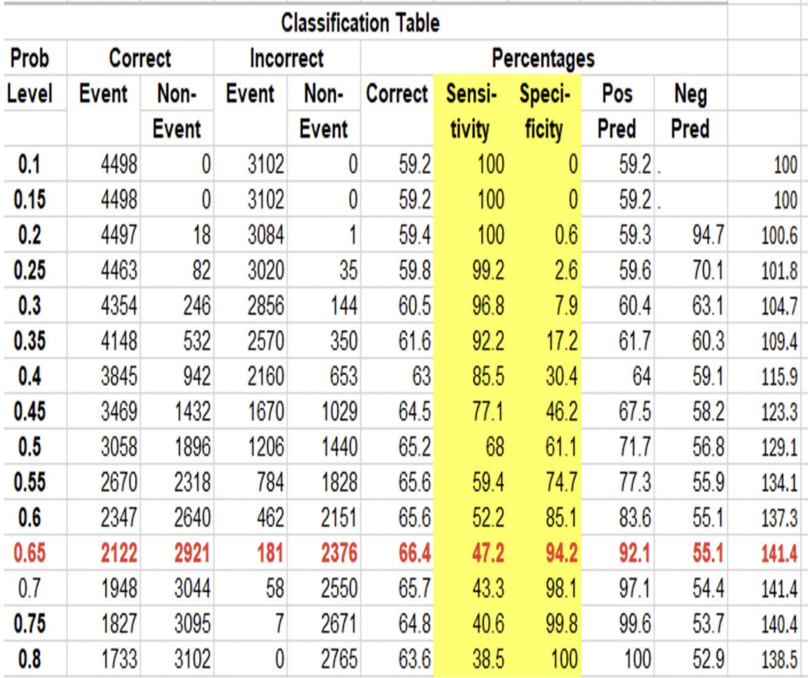
Predicted probabilities and Confidence Intervals

Obs	Customer_care_calls	d_rating_3	Prior_purchases	d_prod_imp_high	Discount_offered	Weight_in_gms	Selected	Warehouse_block	Mode_of_Shipment
1	1	0	0	1	0	0	.		
2	3	0	0	1	0	0	.		
3	1	0	0	0	0	0	.		
4	3	0	0	0	0	0	.		
5	4	0	3	1	29	2602	1	F	Flight
6	4	0	3	1	59	2020	1	A	Flight
7	5	0	4	1	42	1642	1	D	Flight
8	3	1	2	1	7	3311	1	C	Flight
9	2	0	6	1	17	1764	1	B	Flight
10	4	0	4	1	29	1262	0	F	Flight

d_MS_Flight	d_MS_Road	d_rating_1	d_rating_2	d_rating_4	d_rating_5	d_prod_imp_low	d_prod_imp_medium	d_Gender	new_y	_LEVEL_	phat	lcl	ucl
.	.	.	.	.	.	.	.	.	.	1	0.84662	0.79023	0.88996
.	.	.	.	.	.	.	.	.	.	1	0.77627	0.71341	0.82865
.	.	.	.	.	.	.	.	.	.	1	0.79800	0.73949	0.84610
.	.	.	.	.	.	.	.	.	.	1	0.71291	0.65479	0.76476
1	0	0	0	1	0	0	0	1	1	1	0.96563	0.95382	0.97450
1	0	0	0	1	0	0	0	1	1	1	0.99900	0.99819	0.99944
1	0	0	0	0	1	0	0	0	1	1	0.99155	0.98710	0.99447
1	0	0	0	0	0	0	0	1	1	1	0.75221	0.70646	0.79292
1	0	0	0	0	1	0	0	0	1	1	0.91885	0.89602	0.93702
1	0	0	0	0	1	0	0	1	.	1	0.97409	0.96484	0.98095
1	0	0	0	0	1	0	0	1	1	1	0.96874	0.95601	0.97787
1	0	0	0	0	1	0	0	1	1	1	0.99911	0.99832	0.99953

Classification Table

The classification table was generated, to identify the Threshold value.



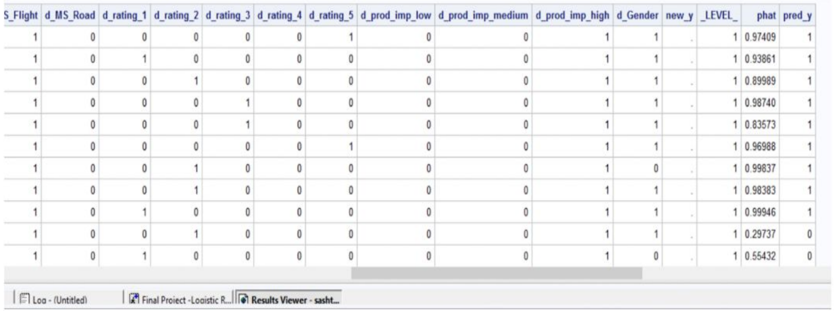
Classification Table

Prob	Correct		Incorrect		Percentages					
Level	Event	Non-	Event	Non-	Correct	Sensi-	Speci-	Pos	Neg	
		Event		Event		tivity	ficity	Pred	Pred	
0.1	4498	0	3102	0	59.2	100	0	59.2	.	100
0.15	4498	0	3102	0	59.2	100	0	59.2	.	100
0.2	4497	18	3084	1	59.4	100	0.6	59.3	94.7	100.6
0.25	4463	82	3020	35	59.8	99.2	2.6	59.6	70.1	101.8
0.3	4354	246	2856	144	60.5	96.8	7.9	60.4	63.1	104.7
0.35	4148	532	2570	350	61.6	92.2	17.2	61.7	60.3	109.4
0.4	3845	942	2160	653	63	85.5	30.4	64	59.1	115.9
0.45	3469	1432	1670	1029	64.5	77.1	46.2	67.5	58.2	123.3
0.5	3058	1896	1206	1440	65.2	68	61.1	71.7	56.8	129.1
0.55	2670	2318	784	1828	65.6	59.4	74.7	77.3	55.9	134.1
0.6	2347	2640	462	2151	65.6	52.2	85.1	83.6	55.1	137.3
0.65	2122	2921	181	2376	66.4	47.2	94.2	92.1	55.1	141.4
0.7	1948	3044	58	2550	65.7	43.3	98.1	97.1	54.4	141.4
0.75	1827	3095	7	2671	64.8	40.6	99.8	99.6	53.7	140.4
0.8	1733	3102	0	2765	63.6	38.5	100	100	52.9	138.5

Confusion Matrix

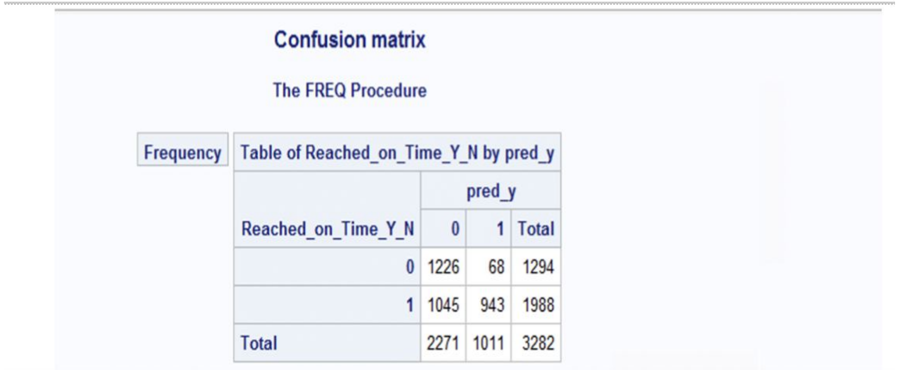
- Predicted probability for the Test set was computed.
- Predicted Y was computed, and the Confusion matrix was generated.

Pred y



S_Flight	d_MS_Road	d_rating_1	d_rating_2	d_rating_3	d_rating_4	d_rating_5	d_prod_imp_low	d_prod_imp_medium	d_prod_imp_high	d_Gender	new_y	_LEVEL_	phat	pred_y
1	0	0	0	0	0	1	0	0	1	1	.	1	0.97409	1
1	0	1	0	0	0	0	0	0	1	1	.	1	0.93861	1
1	0	0	1	0	0	0	0	0	1	1	.	1	0.89989	1
1	0	0	0	1	0	0	0	0	1	1	.	1	0.98740	1
1	0	0	0	1	0	0	0	0	1	1	.	1	0.83573	1
1	0	0	0	0	0	1	0	0	1	1	.	1	0.96988	1
1	0	0	1	0	0	0	0	0	1	0	.	1	0.99837	1
1	0	0	1	0	0	0	0	0	1	1	.	1	0.98383	1
1	0	1	0	0	0	0	0	0	1	1	.	1	0.99946	1
1	0	0	1	0	0	0	0	0	1	1	.	1	0.29737	0
1	0	1	0	0	0	0	0	0	1	0	.	1	0.55432	0

Confusion Matrix



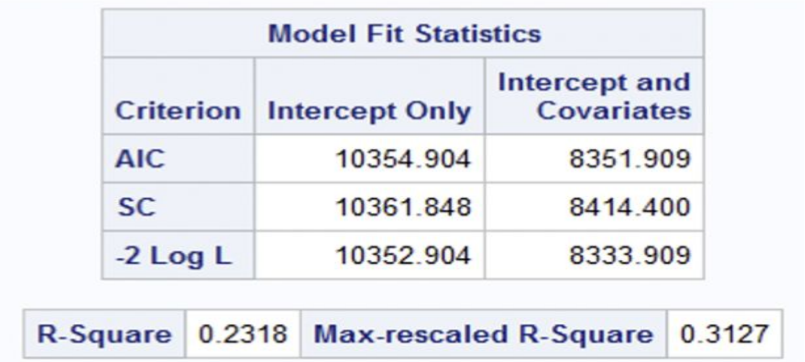
Confusion matrix

The FREQ Procedure

Frequency	Table of Reached_on_Time_Y_N by pred_y			
		pred_y		
	Reached_on_Time_Y_N	0	1	Total
	0	1226	68	1294
	1	1045	943	1988
	Total	2271	1011	3282

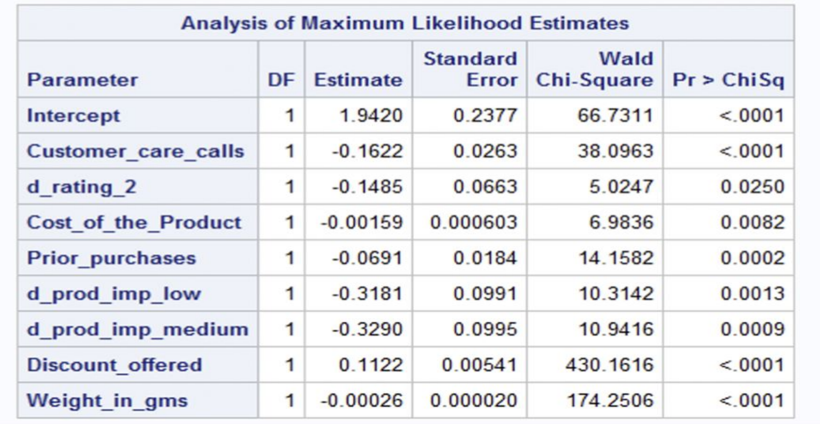
Backward Selection Method

- The same steps were repeated using
- the Backward Selection Method
- The full regression model using Backward selection method resulted in a model with 8 significant predictors, and a R-Square value of 0.23.



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	10354.904	8351.909
SC	10361.848	8414.400
-2 Log L	10352.904	8333.909

R-Square	0.2318	Max-rescaled R-Square	0.3127

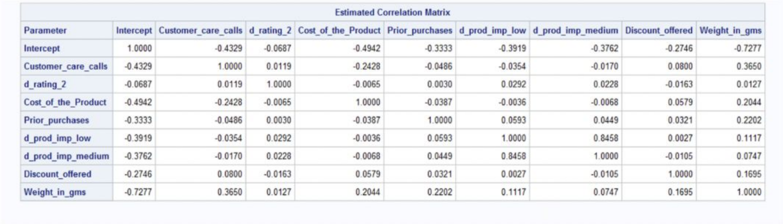


Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.9420	0.2377	66.7311	<.0001
Customer_care_calls	1	-0.1622	0.0263	38.0963	<.0001
d_rating_2	1	-0.1485	0.0663	5.0247	0.0250
Cost_of_the_Product	1	-0.00159	0.000603	6.9836	0.0082
Prior_purchases	1	-0.0691	0.0184	14.1582	0.0002
d_prod_imp_low	1	-0.3181	0.0991	10.3142	0.0013
d_prod_imp_medium	1	-0.3290	0.0995	10.9416	0.0009
Discount_offered	1	0.1122	0.00541	430.1616	<.0001
Weight_in_gms	1	-0.00026	0.000020	174.2506	<.0001

Final Model with Diagnostics

- The final model was run with diagnostics.
- There were no issues of Multicollinearity detected.
- No outliers were detected.
- Several influential points were detected, some were removed, but not much improvement was shown by the model after removing the influential points.

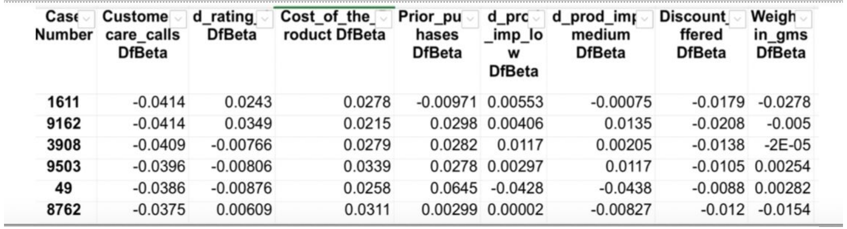
Multicollinearity Scan



Estimated Correlation Matrix

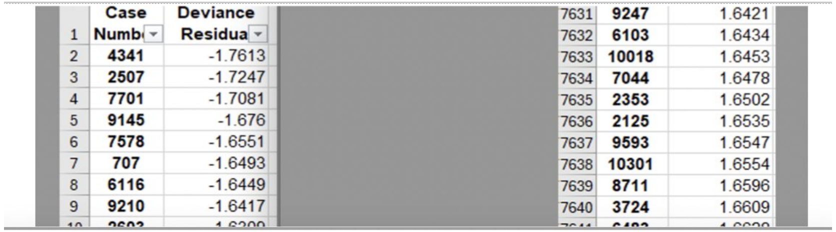
Parameter	Intercept	Customer_care_calls	d_rating_2	Cost_of_the_Product	Prior_purchases	d_prod_imp_low	d_prod_imp_medium	Discount_offered	Weight_in_gms
Intercept	1.0000	-0.4329	-0.0687	-0.4942	-0.3333	-0.3919	-0.3762	-0.2746	-0.7277
Customer_care_calls	-0.4329	1.0000	0.0119	-0.2428	-0.0486	-0.0354	-0.0170	0.0800	0.3650
d_rating_2	-0.0687	0.0119	1.0000	-0.0065	0.0030	0.0292	0.0228	-0.0163	0.0127
Cost_of_the_Product	-0.4942	-0.2428	-0.0065	1.0000	-0.0387	-0.0036	-0.0068	0.0579	0.2044
Prior_purchases	-0.3333	-0.0486	0.0030	-0.0387	1.0000	0.0593	0.0449	0.0321	0.2202
d_prod_imp_low	-0.3919	-0.0354	0.0292	-0.0036	0.0593	1.0000	0.8458	0.0027	0.1117
d_prod_imp_medium	-0.3762	-0.0170	0.0228	-0.0068	0.0449	0.8458	1.0000	-0.0105	0.0747
Discount_offered	-0.2746	0.0800	-0.0163	0.0579	0.0321	0.0027	-0.0105	1.0000	0.1695
Weight_in_gms	-0.7277	0.3650	0.0127	0.2044	0.2202	0.1117	0.0747	0.1695	1.0000

Influential points Scan



Case Number	Custome care_calls DfBeta	d_rating_ DfBeta	Cost_of_the_ roduct DfBeta	Prior_pu hases DfBeta	d_pro _imp_lo w DfBeta	d_prod_imp medium DfBeta	Discount_ ffered DfBeta	Weigh in_gms DfBeta
1611	-0.0414	0.0243	0.0278	-0.00971	0.00553	-0.00075	-0.0179	-0.0278
9162	-0.0414	0.0349	0.0215	0.0298	0.00406	0.0135	-0.0208	-0.005
3908	-0.0409	-0.00766	0.0279	0.0282	0.0117	0.00205	-0.0138	-2E-05
9503	-0.0396	-0.00806	0.0339	0.0278	0.00297	0.0117	-0.0105	0.00254
49	-0.0386	-0.00876	0.0258	0.0645	-0.0428	-0.0438	-0.0088	0.00282
8762	-0.0375	0.00609	0.0311	0.00299	0.00002	-0.00827	-0.012	-0.0154

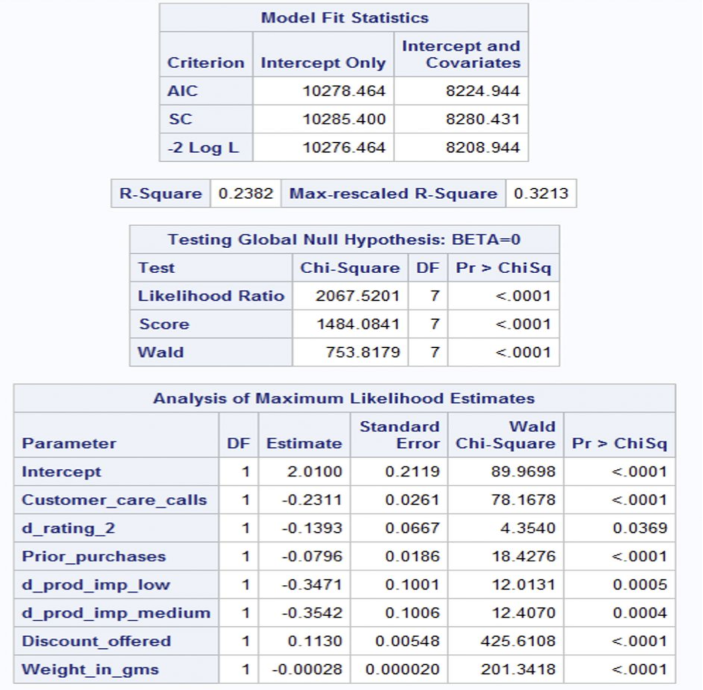
Outliers Scan



	Case	Deviance	7631	9247	1.6421
1	Numb	Residua	7632	6103	1.6434
2	4341	-1.7613	7633	10018	1.6453
3	2507	-1.7247	7634	7044	1.6478
4	7701	-1.7081	7635	2353	1.6502
5	9145	-1.676	7636	2125	1.6535
6	7578	-1.6551	7637	9593	1.6547
7	707	-1.6493	7638	10301	1.6554
8	6116	-1.6449	7639	8711	1.6596
9	9210	-1.6417	7640	3724	1.6609

Final Model

- One insignificant predictor was identified removed.
- The final model had 7 significant predictors, and an R-Square value of 0.2382



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	10278.464	8224.944
SC	10285.400	8280.431
-2 Log L	10276.464	8208.944

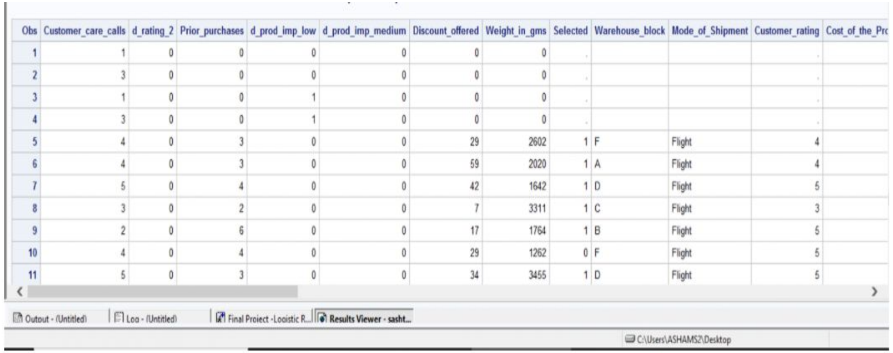
R-Square	0.2382	Max-rescaled R-Square	0.3213

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2067.5201	7	<.0001
Score	1484.0841	7	<.0001
Wald	753.8179	7	<.0001

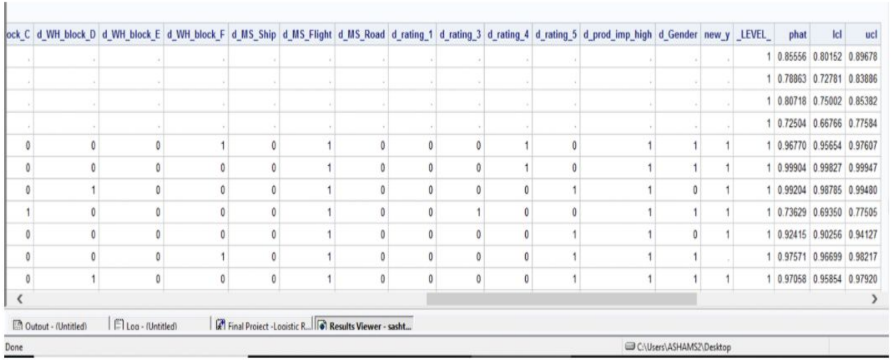
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.0100	0.2119	89.9698	<.0001
Customer_care_calls	1	-0.2311	0.0261	78.1678	<.0001
d_rating_2	1	-0.1393	0.0667	4.3540	0.0369
Prior_purchases	1	-0.0796	0.0186	18.4276	<.0001
d_prod_imp_low	1	-0.3471	0.1001	12.0131	0.0005
d_prod_imp_medium	1	-0.3542	0.1006	12.4070	0.0004
Discount_offered	1	0.1130	0.00548	425.6108	<.0001
Weight_in_gms	1	-0.00028	0.000020	201.3418	<.0001

Predicted Probabilities and Confidence Intervals

- Predictions were computed.
- Datasets were merged.
- Predicted Probabilities, and
- Confidence intervals were generated.
- Phat = 0.856 = 85.6%
- Lcl = [exp(0.80)-1]*100 = 122.55%
- Ucl = [exp(0.89)-1]*100 = 143.51%
- If a shipment has Customer_care_calls = 1, and d_prod_imp_high = 1. The predicted probability of Reached_on_time is 85.6%, it is expected to fall within the range of 122.55% - 143.51% confidence interval.



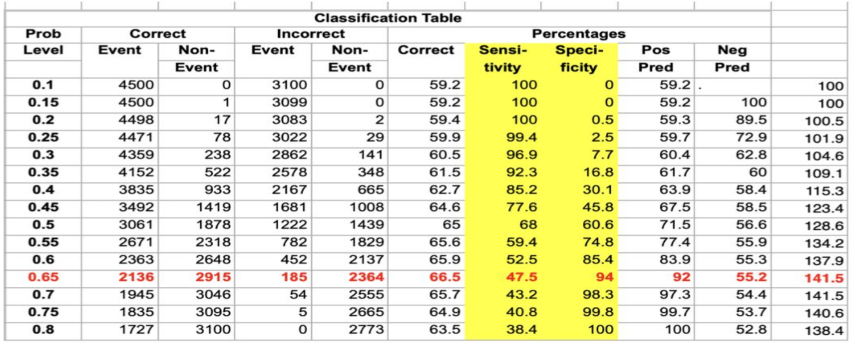
Obs	Customer_care_calls	d_rating_2	Prior_purchases	d_prod_imp_low	d_prod_imp_medium	Discount_offered	Weight_in_gms	Selected	Warehouse_block	Mode_of_Shipment	Customer_rating	Cost_of_the_Pr
1	1	0	0	0	0	0	0	.			.	
2	3	0	0	0	0	0	0	.			.	
3	1	0	0	1	0	0	0	.			.	
4	3	0	0	1	0	0	0	.			.	
5	4	0	3	0	0	29	2602	1	F	Flight	4	
6	4	0	3	0	0	59	2020	1	A	Flight	4	
7	5	0	4	0	0	42	1642	1	D	Flight	5	
8	3	0	2	0	0	7	3311	1	C	Flight	3	
9	2	0	6	0	0	17	1764	1	B	Flight	5	
10	4	0	4	0	0	29	1262	0	F	Flight	5	
11	5	0	3	0	0	34	3455	1	D	Flight	5	



ock_C	d_WH_block_D	d_WH_block_E	d_WH_block_F	d_MS_Ship	d_MS_Flight	d_MS_Road	d_rating_1	d_rating_3	d_rating_4	d_rating_5	d_prod_imp_high	d_Gender	new_y	_LEVEL_	phat	lcl	ucl
.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	0.85556	0.80152	0.89678
.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	0.78863	0.72781	0.83886
.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	0.80718	0.75002	0.85382
.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	0.72504	0.66766	0.77584
0	0	0	1	0	1	0	0	0	1	0	1	1	1	1	0.96770	0.95654	0.97607
0	0	0	0	0	1	0	0	0	1	0	1	1	1	1	0.99904	0.99827	0.99947
0	1	0	0	0	1	0	0	0	0	1	1	0	1	1	0.99204	0.98785	0.99480
1	0	0	0	0	1	0	0	1	0	0	1	1	1	1	0.73629	0.69350	0.77505
0	0	0	0	0	1	0	0	0	0	1	1	0	1	1	0.92415	0.90256	0.94127
0	0	0	1	0	1	0	0	0	0	1	1	1	.	1	0.97571	0.96699	0.98217
0	1	0	0	0	1	0	0	0	0	1	1	1	1	1	0.97058	0.95854	0.97920

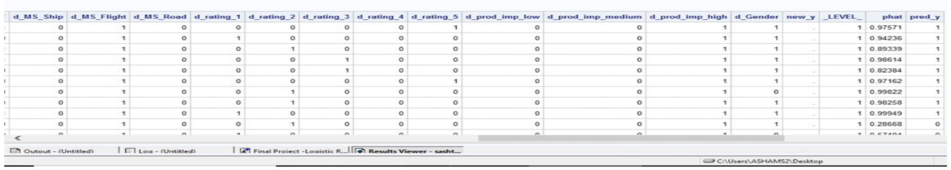
Classification Table and Confusion Matrix

- The classification table was generated, to identify the Threshold value.
- Predicted probability for the Test set was computed.
- Predicted Y was computed, and the Confusion matrix was generated.



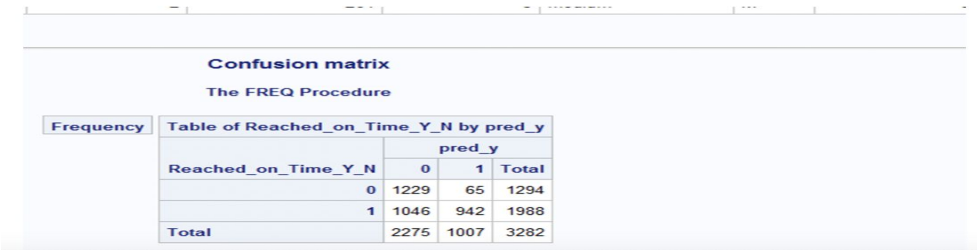
Classification Table

Prob Level	Correct		Incorrect		Percentages					
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	Pos Pred	Neg Pred	
0.1	4500	0	3100	0	59.2	100	0	59.2	.	100
0.15	4500	1	3099	0	59.2	100	0	59.2	100	100
0.2	4498	17	3083	2	59.4	100	0.5	59.3	89.5	100.5
0.25	4471	78	3022	29	59.9	99.4	2.5	59.7	72.9	101.9
0.3	4359	238	2862	141	60.5	96.9	7.7	60.4	62.8	104.6
0.35	4152	522	2578	348	61.5	92.3	16.8	61.7	60	109.1
0.4	3835	933	2167	665	62.7	85.2	30.1	63.9	58.4	115.3
0.45	3492	1419	1681	1008	64.6	77.6	45.8	67.5	58.5	123.4
0.5	3061	1878	1222	1439	65	68	60.6	71.5	56.6	128.6
0.55	2671	2318	782	1829	65.6	59.4	74.8	77.4	55.9	134.2
0.6	2363	2648	452	2137	65.9	52.5	85.4	83.9	55.3	137.9
0.65	2136	2915	185	2364	66.5	47.5	94	92	55.2	141.5
0.7	1945	3046	54	2555	65.7	43.2	98.3	97.3	54.4	141.5
0.75	1835	3095	5	2665	64.9	40.8	99.8	99.7	53.7	140.6
0.8	1727	3100	0	2773	63.5	38.4	100	100	52.8	138.4



d_MS_Ship	d_MS_Flight	d_MS_Road	d_rating_1	d_rating_2	d_rating_3	d_rating_4	d_rating_5	d_prod_imp_low	d_prod_imp_medium	d_prod_imp_high	d_Gender	new_y	_LEVEL_	phat	pred_y
0	1	0	0	0	0	0	1	0	0	1	1	.	1	0.97571	1
0	1	0	1	0	0	0	0	0	0	1	1	.	1	0.94236	1
0	1	0	0	1	0	0	0	0	0	1	1	.	1	0.89339	1
0	1	0	0	0	1	0	0	0	0	1	1	.	1	0.98614	1
0	1	0	0	0	1	0	0	0	0	1	1	.	1	0.82384	1
0	1	0	0	0	0	0	1	0	0	1	1	.	1	0.97162	1
0	1	0	0	1	0	0	0	0	0	1	0	.	1	0.99822	1
0	1	0	0	1	0	0	0	0	0	1	1	.	1	0.98258	1
0	1	0	1	0	0	0	0	0	0	1	1	.	1	0.99949	1
0	1	0	0	1	0	0	0	0	0	1	1	.	1	0.28668	0

Confusion Matrix



Confusion matrix

The FREQ Procedure

Frequency	Table of Reached_on_Time_Y_N by pred_y			
		pred_y		
	Reached_on_Time_Y_N	0	1	Total
	0	1229	65	1294
	1	1046	942	1988
	Total	2275	1007	3282

Model Comparison of Train and Test Performance

Selection Method: **Stepwise**

Sample rate: **70/30**

Seed: **7775559**

Train Performance

X’s in the final model: **Customer_care_calls, d_rating_3, Prior_purchases, d_prod_imp_high, Discount_offered, Weight_in_gms**

R-Square: **23.86**

AIC: **8220.479**

SC: **8269.030**

Selection Method: **Backward**

Sample rate: **70/30**

Seed: **7775559**

Train Performance

X’s in the final model: **Customer_care_calls, d_rating_2, Prior_purchases, d_prod_imp_low, d_prod_imp_medium, Discount_offered, Weight_in_gms**

R-Square: **23.82**

AIC: **8224.944**

SC: **8280.43**

T

Model Comparison of Test Performance

Selection Method: **Stepwise**

Sample rate: **70/30**

Seed: **7775559**

Test Performance

Threshold: **0.65**

TN=1226 FP= 68

FN=1045 TP= 943

Sensitivity : TP/(TP+FN) = **943/(943+1045) =943/1988 = 0.47**

Accuracy :(TP+TN) / (TP+TN+FP+FN) = **(943+1226) / (943+1226+68+1045)= 0.66**

Precision : TP/(TP+FP) = **943/(943+68) = 0.93**

Specificity : TN/(TN+FP) = **1226/(1226+68) =0.95**

election Method: **Backward**

Sample rate: **70/30**

Seed: **7775559**

Test Performance

Threshold: **0.65**

TN= 1229 FP= 65

FN= 1046 TP= 942

Sensitivity = **TP/(TP+FN) = 942/(942+1046) = 0.47**

Accuracy = **(TP+TN) / (TP+TN+FP+FN)= (942+1229)/(942+1229+65+1046)= 0.66**

Precision = **TP/(TP+FP)=942/(942+65)= 0.93**

Specificity = **TN/(TN+FP)=1229/(1229+65)= 0.95**

Best Model

Train performance:

- In terms of training performance the first model with stepwise selection method is slightly better than the other model with backward selection.
- The final has one less predictor, slightly higher R-square value, and slightly lower AIC and SC error terms.

Test performance:

- Both the model have same metrics when it comes to test performance.
- They have the same Threshold, Sensitivity, Accuracy, Precision, and Specificity.

Over all Performance: The first model with the stepwise selection method is slightly better than the second model .