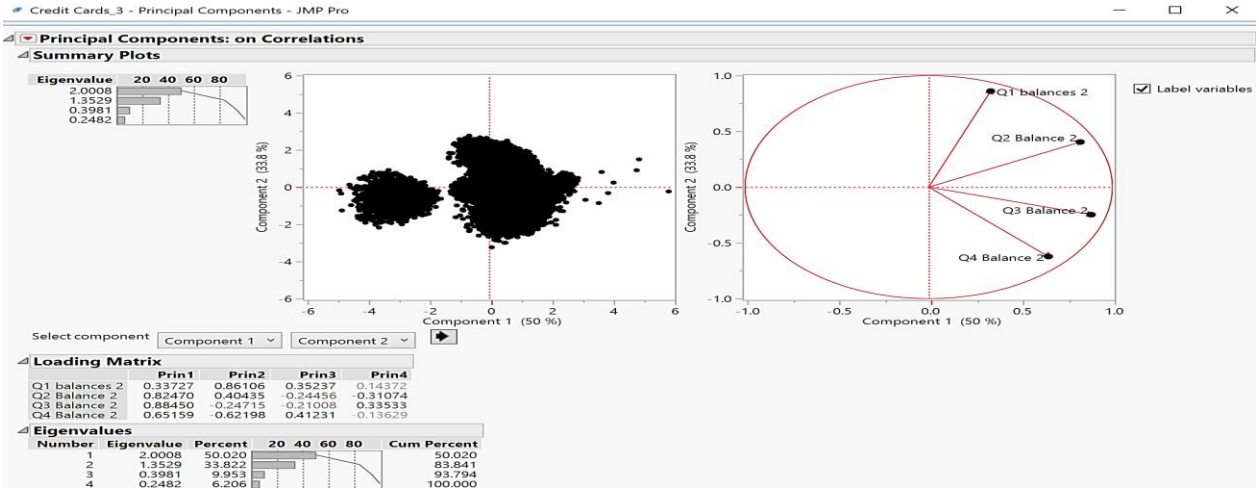
Building a Logistic Regression model to predict customer's response to a credit card offer

Exploratory data analysis

- As Credit Rating increases, the number of responses decrease. Likewise, as the variables Homes Owned and Bank Accounts Open increase, the number of responses decline.
- As Q1,Q2,Q3,Q4 Balances increase, the number of responses decrease.
- Customers with no Overdraft protection responded more than those who didn't have protection.
- Customers who owned their home responded more than those who didn't own.
- Reward type 'Air miles' has more responses than Points and Cashback.
- No definite relation with response variable has been found for variables Credit Cards Held, Household Size and Income Level.

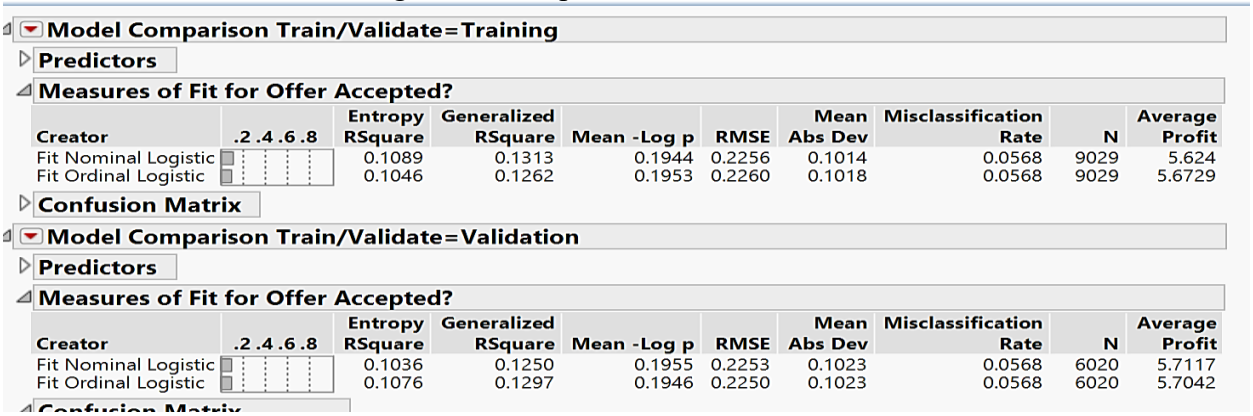
No significant correlation was found among predictor variables. Slight correlation was noticed between Q1-Q4 Balance variables. Thus we considered these variables for Principal Component Analysis.

Dimension Reduction



Dimension reduction is performed by taking only first 3 principal components into the model with 93.794% information. This ensures that there is less loss of data and the data retained conveys the same information concisely.

Here is the screen shot showing model comparison between PCA and without PCA.



Model Comparison Train/Validate=Training

Predictors

Measures of Fit for Offer Accepted?

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N	Average Profit
Fit Nominal Logistic		0.1089	0.1313	0.1944	0.2256	0.1014	0.0568	9029	5.624
Fit Ordinal Logistic		0.1046	0.1262	0.1953	0.2260	0.1018	0.0568	9029	5.6729

Confusion Matrix

Model Comparison Train/Validate=Validation

Predictors

Measures of Fit for Offer Accepted?

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N	Average Profit
Fit Nominal Logistic		0.1036	0.1250	0.1955	0.2253	0.1023	0.0568	6020	5.7117
Fit Ordinal Logistic		0.1076	0.1297	0.1946	0.2250	0.1023	0.0568	6020	5.7042

Confusion Matrix

Since the Entropy R squared is highest for this model and this is providing Average profit more than baseline models this model was chosen.

How good is the model?

Since I have data to create profit matrix, following "Decision count" matrix is created to help evaluate how good my model is. The profit matrix is

Specified Profit Matrix

	Decision	
Actual	**No**	**Yes**
No	0	-2.25
Yes	-138.5	136.25

Frequencies

Level	Count	Prob
No	5678	0.94319
Yes	342	0.05681
Total	6020	1.00000

This is used on all the baseline models as well as the chosen model to find how far the model is profitable.

Base Line Model 1: Contact No One

From the validation data in the given Table I found there are 5678 – "No" and 342 – "Yes" When this value is used along with above profit matrix, the Estimated Net profit for contacting No One is: $0(5678) + (-138.5) * (342) = -47,367

Per Customer: -47,367 / (5678 + 342) = -$7.862

Base Line Model 2: Contact Every One

When same profit matrix and same frequencies are used from above, when we wanted to contact everyone, the Estimated profit is: -2.25 * (5678) + 136.25 * (342) = 33,822

So Estimated profit per customer is: 33,822 / (5678 + 342) = +$5.618

Model 3: The chosen model with PCA

The model has the following confusion matrix before considering any profit values:

Actual Offer Accepted?	Predicted Count No	Yes
No	5678	0
Yes	342	0

Actual Offer Accepted?	Decision Count No	Yes
No	352	5326
Yes	1	341

The average profit per customer is +$ 5.7042. This shows the model chosen is better than baseline models.

Model Performance Summary

	Entropy R square Train	Entropy R square Validation	Net Profit per Customer
Logistic Regression	10.46%	10.76%	**$5.7042**
Base Line Contact No One			**-$7.862**
Base Line Contact Every One			**+$ 5.618**