Assignment 2: Discriminant Analysis

Rajwinder Mahal, Shandhra Ramana, Shayan Toor, Pratheek Kumar

BANA 271: Marketing Analytics

Prof. Rajeev Tyagi

February 6, 2022

Optimal Model and Relative Importance of Factors

To avoid over prediction, we first divided our data into a training and testing set. We randomly assigned 70% of the data for training and the remaining 30% for testing. We also used leave one-out classification. Subsequently, a consumer will either own a tablet or not. Based on the given data, there are 22 respondents who own a tablet and 50 respondents who don't. So, if we say that everyone owns a tablet, then we will be correct 30.5% of the time. On the other hand, if we say that everyone doesn't own a tablet, then we will be correct 69.44% of the time. Given this information, our model should perform better than 69.44%. But if we take the average of both situations, then our model should perform better than 49.97%.

To find the relative importance of factors, we decided to go with an experimental model with all the variables. We are calling it Model 1.1 and the SPSS outputs for this model are shown in the Appendix Figure 2, 3, 4, and 5. This model has a cross-validated accuracy of 74.5% which is higher than the threshold accuracy. Also, this model has a p-value of 0.003 which is less than 0.05 and thus, this model is explaining statistically significant amounts of consumer behavior in regards to whether they own a tablet or not. Moreover, appendix figure 2 shows the relative importance of each variable/factor. Based on this output, we found that income and innovators are the most important factors with 22% and 21% relative importance, respectively. Scheduler and remote_meet are the next in the line with relative importance of 17% and 14%, respectively. Infoshare and time_info have the least influence on the model with relative importance of 0% and 1%, respectively. Based on this information, we think that anyone who has a high income and is open to trying new technologies in their field should own a tablet. For example, if we take an example from the real world, most people with high income would most likely own a tablet.

Although we have a good model that meets both p-value and accuracy thresholds, we would like to have a model that has good accuracy with a minimum number of variables. For example, if we are in a situation where it is important to make a good prediction in the least amount of time, then we should find the next best model that has similar accuracy with fewer variables. Moreover, there will be some consumers who wouldn't be willing to give too much information. So, based on the relative importance of factors from Model 1.1, we came up with 11 different models as shown in appendix figure 1. Based on the eigenvalue, p-value, and prediction accuracy, we think Model 1.7 is the optimal model for determining whether a consumer owns a computing tablet. The SPSS outputs for Model 1.7 are shown in appendix figure 9, 10, and 11. This model uses Innovator, Scheduler, Remote_meet, and Income variables. Apart from the relative importance of these factors, we think these are usually the factors that determine whether someone has a tablet or not in the real world. People who have the money and do some kind of scheduling or remote work are more likely to own a tablet than someone with low income. The accuracy of this model is 78.4% which is higher than that of Model 1.1 with 10 variables. Generally, adding additional independent variables to the model improves the accuracy. But in our case, the accuracy gets worse if we go with 10 variables instead of 4 variables used in Model 1.7. We think this might be because of some unknown factors such as independent variables not being completely independent in the real world. Also, this model has a p-value of less than 0.001 which is even lower than the p-value of 0.003 for Model 1.1. This means this model explains a statistically significant amount of grouping as compared to Model 1.1. Moreover, this model is using less parameters while providing the same accuracy and thus, we think this is the optimal model. Finally, for Model 1.7, Innovator and Income has the greatest relative importance of 28% each. Remote_meet and Scheduler have relative importance of 24% and 20%. Given these

numbers, we can conclude that all variables in this model have almost the same relative importance.

Model if data is freely available

When data is freely available, we should use as many variables as possible because adding extra independent variables doesn't have a negative effect on the model accuracy. Adding them can only increase the model accuracy. Also, we are assuming that asking consumers additional questions will not have any indirect costs such as privacy issues, consumers thinking the company is collecting too much personal information, etc. Given these assumptions, we think Model 1.3 is the optimal model for determining whether a consumer owns a computing tablet if the data is freely available. The SPSS outputs for Model 1.3 are shown in appendix figure 6, 7, and 8. This model uses Innovator, Slack, Scheduler, Work, Remote_meet, Age, Education, and Income variables. The accuracy of this model is 74.5% which is the same as Model 1.1. However, this model has a p-value of less than 0.001 which is even lower than the p-value of 0.003 for Model 1.1. This means this model explains a statistically significant amount of grouping as compared to Model 1.1. Moreover, this model is using fewer parameters while providing the same accuracy and thus, we think this is the optimal model to use if data is freely available. We could have used Model 1.1 which has 10 variables, however, the relative importance of info_share and time_info was 0% and 1%, respectively. Because of this very low relative importance, we decided to exclude those variables from our model because there is always some cost associated with data whether direct or indirect. At the same time, if we think there is no cost associated with it, then there is always processing time that we must take into account. Adding extra parameters would add time to model prediction and nowadays it is very important to have the lowest latency and there are all types of cloud computing solutions for edge computing to lower the latency. So, taking all this information into account, we decided to go with Model 1.3. Moreover, Model 1.3 has the same prediction accuracy as that of Model 1.1. Finally, for Model 1.3, Income and Innovator have the greatest relative importance with 23% and 21% respectively. Work and age has the least relative importance, 3% and 5%, respectively.

Model if data is costly to collect

When data collection is very costly, then we should have a model that has better prediction accuracy than the threshold while using the minimum number of variables. Given this, we found that Model 1.9 (shown in appendix figure 12, 13, and 14) will be the best model. It only uses two variables, Income and Innovator and has prediction accuracy of 80.4% which is higher than what we have seen in the previous models. It is possible that our model is overfitting, however, it has 71.4% accuracy on unseen data. This means our model is not overfitting the data and hence is good to use in the situation where data collection is very costly. In this model, income has slightly higher importance of 59% as compared to Innovators which has relative importance of 41%. Moreover, we found that if we only use the income variable (Model 2.1, appendix figure 1), then it has 78.4% accuracy. However, we chose to go with Model 1.9 with 2 variables to avoid any unknown consequences such as overly depending on income factor.

Conclusion

Based on different models with different numbers of variables/factors, we found out that income and innovators have the greatest relative importance in every single model we came up with. Removing them significantly lowers the prediction accuracy. Ultimately, we think Model 1.7 with 4 variables (income, innovators, scheduler, and remote_meet) is the optimal model.

Appendix

Figure 1: **Summary of Models**

Model Version	Number of Variables	Factors	Eigenvalue	Wilks' Lambda Sig.	Cross Validated	Model Accuracy (cases not selected)
1.1	10	Innovator, Slack, Scheduler, Time_Info, Work, Remote_meet, Infoshare, Age, Education, Income	0.862	0.003	74.5%	76.2%
1.2	9	Innovator, Slack, Scheduler, Time_Info, Work, Remote_meet, Age, Education, Income	0.826	0.002	74.5%	76.2%
1.3	**8**	**Innovator, Slack, Scheduler, Work, Remote_meet, Age, Education, Income**	**0.825**	**< 0.001**	**74.5%**	**76.2%**
1.4	7	Innovator, Slack, Scheduler, Remote_meet, Age, Education, Income	0.818	< 0.001	74.5%	71.4%
1.5	6	Innovator, Slack, Scheduler, Remote_meet, Education, Income	0.791	< 0.001	78.4%	71.4%
1.6	5	Innovator, Scheduler, Remote_meet, Education, Income	0.767	< 0.001	76.5%	71.4%
1.7	**4**	**Innovator, Scheduler, Remote_meet, Income**	**0.730**	**< 0.001**	**78.4%**	**76.2%**
1.8	3	Innovator, Scheduler, Income	0.565	< 0.001	78.4%	76.2%
1.9	**2**	**Innovator, Income**	**0.542**	**< 0.001**	**80.4%**	**71.4%**
2.1	1	Income	0.400	< 0.001	78.4%	71.4%
2.2	2	Innovator, Scheduler	0.252	0.005	72.5%	66.7%

Figure 2: **Model 1.1**

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.826[a]	100.0	100.0	.673

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.548	26.486	10	.003

Standardized Canonical Discriminant Function Coefficients

	Function 1
Innovator	.680
Income	.723
Scheduler	-.537
work	-.105
Infoshare	.009
Education	-.284
Age	-.176
remote meet	-.457
Slack	.220
Time_Info	-.025

Figure 3: **Model 1.1**

Stepwise Statistics

Variables Entered/Removed[a,b,c,d]

Step	Entered	Wilks' Lambda Statistic	df1	df2	df3	Exact F Statistic	df1	df2	Sig.
1	Income	.714	1	1	49.000	19.591	1	49.000	.000
2	Innovator	.648	2	1	49.000	13.010	2	48.000	.000
3	remote meet	.617	3	1	49.000	9.734	3	47.000	.000
4	Scheduler	.578	4	1	49.000	8.393	4	46.000	.000
5	Education	.566	5	1	49.000	6.899	5	45.000	.000
6	Age	.556	6	1	49.000	5.861	6	44.000	.000
7	Slack	.550	7	1	49.000	5.025	7	43.000	.000
8	work	.548	8	1	49.000	4.333	8	42.000	.001
9	Time_Info	.548	9	1	49.000	3.761	9	41.000	.002
10	Infoshare	.548	10	1	49.000	3.303	10	40.000	.003

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

a. Maximum number of steps is 20.

b. Maximum significance of F to enter is 0.99.

c. Minimum significance of F to remove is 1.

d. F level, tolerance, or VIN insufficient for further computation.

Figure 4: **Model 1.1**

Canonical Discriminant Function Coefficients

	Function 1
Innovator	.436
Income	.041
Scheduler	-.291
work	-.096
Infoshare	.007
Education	-.276
Age	-.016
remote meet	-.233
Slack	.143
Time_Info	-.015
(Constant)	-.441

Unstandardized coefficients

Classification Function Coefficients

	Own_Tablet 0	Own_Tablet 1
Innovator	1.722	2.574
Income	.032	.112
Scheduler	1.014	.445
work	3.712	3.525
Infoshare	3.203	3.218
Education	-.364	-.902
Age	.300	.269
remote meet	2.929	2.474
Slack	4.527	4.806
Time_Info	.747	.717
(Constant)	-39.578	-41.226

Fisher's linear discriminant functions

Figure 5: **Model 1.1**

Classification Results[a,b,d]

				Predicted Group Membership		
			Own_Tablet	0	1	Total
Cases Selected	Original	Count	0	32	4	36
			1	5	10	15
		%	0	88.9	11.1	100.0
			1	33.3	66.7	100.0
	Cross-validated[c]	Count	0	30	6	36
			1	7	8	15
		%	0	83.3	16.7	100.0
			1	46.7	53.3	100.0
Cases Not Selected	Original	Count	0	11	3	14
			1	2	5	7
		%	0	78.6	21.4	100.0
			1	28.6	71.4	100.0

a. 82.4% of selected original grouped cases correctly classified.

b. 76.2% of unselected original grouped cases correctly classified.

c. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

d. 74.5% of selected cross-validated grouped cases correctly classified.

Figure 6: **Model 1.3**

Canonical Discriminant Function Coefficients

	Function 1
Innovator	.434
Income	.041
Scheduler	-.293
work	-.096
Education	-.273
Age	-.016
remote meet	-.237
Slack	.142
(Constant)	-.456

Unstandardized coefficients

Classification Function Coefficients

	Own_Tablet 0	Own_Tablet 1
Innovator	.784	1.632
Income	-.011	.068
Scheduler	1.553	.980
work	4.589	4.402
Education	-.430	-.962
Age	.247	.216
remote meet	2.962	2.499
Slack	3.891	4.168
(Constant)	-31.141	-32.818

Fisher's linear discriminant functions

Figure 7: **Model 1.3**

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.825[a]	100.0	100.0	.672

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.548	27.077	8	<.001

Standardized Canonical Discriminant Function Coefficients

	Function 1
Innovator	.677
Income	.723
Scheduler	-.540
work	-.105
Education	-.281
Age	-.175
remote meet	-.465
Slack	.219

Figure 8: Model 1.3

Classification Results[a,b,d]

			Own_Tablet	Predicted Group Membership 0	1	Total
Cases Selected	Original	Count	0	32	4	36
			1	5	10	15
		%	0	88.9	11.1	100.0
			1	33.3	66.7	100.0
	Cross-validated[c]	Count	0	30	6	36
			1	7	8	15
		%	0	83.3	16.7	100.0
			1	46.7	53.3	100.0
Cases Not Selected	Original	Count	0	11	3	14
			1	2	5	7
		%	0	78.6	21.4	100.0
			1	28.6	71.4	100.0

a. 82.4% of selected original grouped cases correctly classified.

b. 76.2% of unselected original grouped cases correctly classified.

c. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

d. 74.5% of selected cross-validated grouped cases correctly classified.

Figure 9: **Model 1.7**

Canonical Discriminant Function Coefficients

	Function 1
Innovator	.415
Income	.037
Scheduler	-.258
remote meet	-.279
(Constant)	-1.366

Unstandardized coefficients

Classification Function Coefficients

	Own_Tablet 0	1
Innovator	.300	1.063
Income	.103	.170
Scheduler	1.583	1.109
remote meet	1.647	1.134
(Constant)	-9.075	-12.280

Fisher's linear discriminant functions

Figure 10: **Model 1.7**

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.730[a]	100.0	100.0	.650

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.578	25.756	4	<.001

Standardized Canonical Discriminant Function Coefficients

	Function 1
Innovator	.648
Income	.654
Scheduler	-.476
remote meet	-.548

Figure 11: **Model 1.7**

Classification Results[a,b,d]

			Own_Tablet	Predicted Group Membership 0	1	Total
Cases Selected	Original	Count	0	31	5	36
			1	6	9	15
		%	0	86.1	13.9	100.0
			1	40.0	60.0	100.0
	Cross-validated[c]	Count	0	31	5	36
			1	6	9	15
		%	0	86.1	13.9	100.0
			1	40.0	60.0	100.0
Cases Not Selected	Original	Count	0	12	2	14
			1	3	4	7
		%	0	85.7	14.3	100.0
			1	42.9	57.1	100.0

a. 78.4% of selected original grouped cases correctly classified.

b. 76.2% of unselected original grouped cases correctly classified.

c. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

d. 78.4% of selected cross-validated grouped cases correctly classified.

Figure 12: **Model 1.9**

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.542[a]	100.0	100.0	.593

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.648	20.790	2	<.001

Standardized Canonical Discriminant Function Coefficients

	Function 1
Innovator	.523
Income	.752

Figure 13: **Model 1.9**

Canonical Discriminant Function Coefficients

	Function 1
Innovator	.335
Income	.042
(Constant)	-3.206

Unstandardized coefficients

Classification Function Coefficients

	Own_Tablet 0	Own_Tablet 1
Innovator	1.066	1.597
Income	.106	.173
(Constant)	-4.482	-10.077

Fisher's linear discriminant functions

Figure 14: **Model 1.9**

Classification Results[a,b,d]

			Own_Tablet	Predicted Group Membership 0	1	Total
Cases Selected	Original	Count	0	32	4	36
			1	6	9	15
		%	0	88.9	11.1	100.0
			1	40.0	60.0	100.0
	Cross-validated[c]	Count	0	32	4	36
			1	6	9	15
		%	0	88.9	11.1	100.0
			1	40.0	60.0	100.0
Cases Not Selected	Original	Count	0	12	2	14
			1	4	3	7
		%	0	85.7	14.3	100.0
			1	57.1	42.9	100.0

a. 80.4% of selected original grouped cases correctly classified.

b. 71.4% of unselected original grouped cases correctly classified.

c. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

d. 80.4% of selected cross-validated grouped cases correctly classified.

Figure 15: Model 2.1

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.400[a]	100.0	100.0	.534

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.714	16.312	1	<.001

Standardized Canonical Discriminant Function Coefficients

	Function 1
Income	1.000

Figure 16: Model 2.1

Canonical Discriminant Function Coefficients

	Function 1
Income	.056
(Constant)	-2.619

Unstandardized coefficients

Classification Function Coefficients

	Own_Tablet 0	Own_Tablet 1
Income	.125	.201
(Constant)	-3.156	-7.099

Fisher's linear discriminant functions

Figure 17: Model 2.1

Classification Results[a,b,d]

				Predicted Group Membership 0	Predicted Group Membership 1	Total
			Own_Tablet			
Cases Selected	Original	Count	0	33	3	36
			1	8	7	15
		%	0	91.7	8.3	100.0
			1	53.3	46.7	100.0
	Cross-validated[c]	Count	0	33	3	36
			1	8	7	15
		%	0	91.7	8.3	100.0
			1	53.3	46.7	100.0
Cases Not Selected	Original	Count	0	12	2	14
			1	4	3	7
		%	0	85.7	14.3	100.0
			1	57.1	42.9	100.0

a. 78.4% of selected original grouped cases correctly classified.

b. 71.4% of unselected original grouped cases correctly classified.

c. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

d. 78.4% of selected cross-validated grouped cases correctly classified.