May 8, 2019 - Team 03 (Mackenzie O., Madison R., Madeline S.)

Analysis of Black Friday Shopping Habits

Abstract

In this analysis, we explore the trends of Black Friday shopping. Using an array of machine learning algorithms, we determined that Product_Category_1 is the most important predictor. In our research, we analysis the population data set and the age range of 18-25, the young millennials. Our goal is to derive attributes that depict characteristics of a high spender, someone who spends more than the average for the population. Through our analysis, the Bayesian Network performed the best and it was found that the range 26-35 actually spends the most, not those from 18-25 like we had hypothesized.

Introduction

We decided to explore the dataset "Black Friday" from the Kaggle website. Black Friday is the day after Thanksgiving. It is considered the first day of Christmas shopping. On Black Friday, stores and malls are usually open all night and have very good deals. It is a time when a lot of people can get their Christmas shopping done without spending too much on each person you are buying for. The Black Friday dataset from Kaggle includes information about the shopping and purchasing habits of Black Friday shoppers. Our dataset analyzes many attributes, including user ID, product ID, gender, age, occupation, city, number of years in current city, marital status, 3 different products and their categories, and purchase amount. Our goal through this research was to explore which age group was likely to spend and buy the most. We also did a separate analysis of those in the millennial age group to see if students/young adults, like us, tend to spend or buy more.

We analyzed both the full population and just those that fall into the millennial age range, 18-25 years old, that are single. First, for both sections we derived "High Spender" as a flag variable, which allowed us to find which age group spent and bought the most. Also, for both sections, we filtered the data to only look those who bought three products. This helped us decrease the amount of data we were dealing with without skewing the results heavily. For the population section, we did four types of analysis. We used CART, Apriori, Neural Net, and Bayes Net. Since we discovered that Bayes Net was the best for the population, we also used it for the millennials section. We also used K Means for the millennials section. Our results can be found below.

We found it pretty easy to find information on our topic. Although it isn't a scholarly topic, it is a well-known "holiday" and there is a lot of information about it. The dataset from Kaggle has plenty of information alone. Unfortunately, even though there is a lot of information on Black Friday, it isn't necessarily the information we were looking for. We had a hard time finding statistical information to support our findings.

Literature Review

Before starting our analysis, we looked into what had already been done in this field of research. At Eastern Illinois University, researchers conducted experiments using different types of observations showing that females were happier than males when shopping. They observed that people lined up for hours outside the stores and that chaos lasted approximately 30 minutes after the stores opened and that the shoppers were 72% female (Simpson, et al).

The website stackabuse published an article about Black Friday trends using the same data set as us (Guest Contributor, Analysis of Black Friday Shopping Trends via Machine Learning). They then plot male vs female, showing that there are almost 3x as many males in the set than females, and the age range of 26-35 contains the most samples. In agreement with this, a *blackfriday.com* study shows that older millennials, aged 25-34 spent the most per person (Staff). This makes sense because people are beginning to become financially stable at that age range. In further analysis, *stackabuse* gathers the average amount spent for each occupation. They then used machine learning to predict the amount spent for each using linear regression. This is most similar to ours, except we predicted high-spender and used multiple ML algorithms.

According to *finder.com*, people who are married or in a domestic partnership are 78% more likely than other relationship statuses to spend money on Black Friday (McDermott). Those who have never married are next most likely at 73%. Additionally,77% of men are likely to shop, while 71% of women are likely to shop. Men are expected to spend double that of women, with an average $626.44 versus $342.50. According to *thebalance.com*, traffic in retail stores is declining due to online shopping and Cyber Monday (Amadeo). In fact, in 2018, traffic in stores fell as much as 9%.

Methodology:

In order to analyze our data, we first did a data audit on the full dataset. This helped us to look for missing or null data. There was some null values for those who did not buy three products. Thus, we narrowed our data down to only those who bought three products, not one or two. In order to do this, we used the generate node to filter the data. We found that our dataset was now complete. From there, we used the derive node to create the variable "High Spender." We used this variable to find which age range and marital status spent and bought the most. Next, we partitioned the data. We partitioned it 70% training and 30% testing, because it is the standard way of partitioning. We used the type node to ensure that all the data in the dataset was of the correct type. This type node made sure that the "marital status" variable was set up as a flag and that the "high spender" variable was displayed as either a 0, meaning they were not a "high spender", or a 1, meaning that they were a "high spender."

Next, we worked with the full population and ran four analysis tests on it. These four include CART, Apriori, Neural Net and Bayes Net. After generating the results, we used different performance and evaluation metrics to see which analysis test best represented our data. After examining the full population, we used a filter node again to only work with the data that represented those that were 18 to 25 years old and were single. We called this our millennial section. For the millennial section, we used Bayes Net and K Means. We decided on Bayes Net, not CART, Apriori, or Neural Net, because it was the best and most accurate for the population and therefore would be the best and most accurate for the millennials.

Analysis

In our study we pursued two analyses: an analysis to derive characteristics of a high spender in the entire dataset, and a sub-analysis of unmarried people between the ages of 18-25.

Population Analysis:

Cart Tree (Appendix A)

The prediction algorithm generated results with an 78.5% accuracy for training and testing sets. They have 73.5% precision for each. In Appendix A, you can see a pictorial representation of the decision tree model. The model predicted that Product_Category_1, followed by Product_Category_3, are by far the most important predictors in determining whether an auction is competitive or not. Looking at the rules, it used Product_Category_1 and Product_Category_3 in its model decisions.

Bayes Net (Appendix B)

The model had performance metrics of 90.4% and 90.1% accuracy respectively for training and testing. The models had 94.2% and 93.8% recall, 13.8% and 13.9% FP rate, and 88.2% and 88.1% precision. The accuracy, recall, and precision are decent, but the FP rate is relatively high. An ROC curve was produced to further evaluate the performance of the Bayesian Network. The area under the curve is a measure of the predictive accuracy of the model. In this case, the area under the curve appears to be about 0.95, which is relatively accurate. Looking at the conditional probabilities on age, we see that the highest probability for high spenders is always in the 26-35 years old range.

Neural Net (Appendix C)

Neural networks predict a continuous or categorical target based on one or more predictors by finding unknown and possibly complex patterns in the data. We predicted that the training dataset will be competitive 78.42% of the time and the testing set 78.41%. The models have 78.4% accuracy for both and 73.2% and 73.3% precision in their prediction respectively. We also know that Product_Category_1 is the most important variable in determining if a person is a high spender.

Apriori Modelling (Appendix D)

The network automatically generated 28 association rules, all of which with the same end goal: to predict a high spender. We can determine if each rule is a good or poor rule based on calculating improvement/lift. According to the generated rules, the two best predictors are product_Category_1= 1.0 and City_Category=C based on their support, confidence and lift.

Millennial Analysis:

Bayesian Network (Appendix E)

The model had performance metrics of 89.9% and 89.5% accuracy respectively for training and testing. The models had 93.5% and 93.2% recall, 7.4% and 14.4% FP rate, and 88.0% and 87.3% precision. The accuracy, recall, and precision are decent, but the FP rate is a little high. An ROC curve was produced to further evaluate the performance of the Bayesian Network. The area under the curve is a measure of the predictive accuracy of the model. In this case, the area under the curve appears to be about 0.90, which is relatively accurate. Looking at the conditional probabilities of gender, the millennial high spenders are most likely to be male and the millennial not high spenders are most likely to be male as well. In general, in the millennial age group at least, the majority of Black Friday shoppers are males.

KMeans (Appendix F)

Cluster 5 – comprises 19.4% of the data with 4767 people in the cluster
Cluster 2 - comprises 19.0% of the data with 4672 people in the cluster
Cluster 3 - comprises 15.8% of the data with 3887 people in the cluster
Cluster 1 - comprises 12.0% of the data with 2933 people in the cluster
Cluster 4 - comprises 10.0% of the data with 2449 people in the cluster
Cluster 6 - comprises 8.3% of the data with 2043 people in the cluster
Cluster 7 - comprises 8.1% of the data with 1997 people in the cluster
Cluster 8 - comprises 7.3% of the data with 1793 people in the cluster

The variable highSpender is the most important in clustering people. Cluster-5, cluster-2, and cluster-4 are 100% high spenders. Cluster-3, cluster-6, cluster-7, and cluster-8 are 100% not high spenders. Cluster-1 is 69.0% high spenders. In 7 out of 8 of the clusters, the people are all high spenders or all not high spenders. However, overall, the cluster quality is in the "poor" range so K-Means is not the best method for clustering people.

Results and Conclusions

For the population analysis, the best model for predicting if someone was a high spender or not was the Bayesian Network. We found that the highest probability for high spenders was frequently in the 26-35 years old range, not in the millennial age range like expected. For the millennial analysis, the best model for predicting if someone was a high spender or not was also the Bayesian Network. Through our multiple analyses, the attribute of "Product_Category_1" emerged in almost every model as the most important predictor in determining if a customer is a "high-spender" or not. The apriori model contained results that were mostly inconclusive in exact attributes of what products correlate highly with high spenders even though there was a high volume of data.

For the millennial analysis, we again used a Bayesian network, as it proved to have the best metrics on the population set. In conjunction, we also used K-Means clustering to determine if certain clusters were more likely to be high spenders. In cluster 7, consisting of 100% males from city A, the cluster consisted of 100% high spenders. In the other clusters, many of them had high spender as either 100% 0 or 100% 1, suggesting high correlation between being a high spender and other attributes.

It seems that males dominated our sample, and that people 26-35 were more likely to be high-spenders, so companies should aim their targets at that sample rather than at the young or the old. This makes sense because people are more financially stable at the 26-35 year age range, but have more room to take risks and spend money on themselves.

Sources

Amadeo, Kimberly. “How Much Do Americans Spend on Black Friday?” The Balance, The Balance, 20 Jan. 2019, www.thebalance.com/what-is-black-friday-3305710.

Contributor, Guest. “Analysis of Black Friday Shopping Trends via Machine Learning.” Stack Abuse, Stack Abuse, 1 May 2019, stackabuse.com/analysis-of-black-friday-shopping-trends-via-machine-learning/.

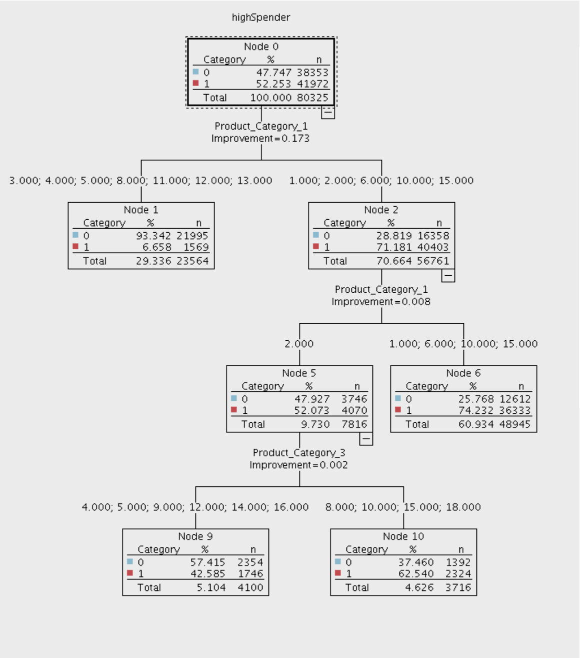
McDermott, Jennifer. “Black Friday Spending Statistics 2018.” Finder US, 6 May 2019, www.finder.com/black-friday-statistics.

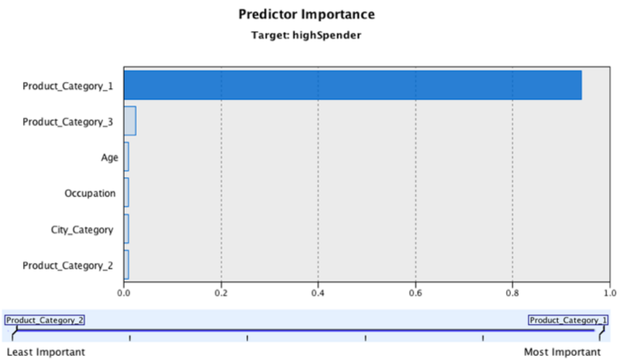
Simpson, Linda, et al. “An Analysis of Consumer Behavior on Black Friday.” Thekeep.eiu.edu, Eastern Illinois University, Eastern Illinois University, July 2011, thekeep.eiu.edu/cgi/viewcontent.cgi?article=1012&context=fcs_fac.

Staff, BlackFriday.com. “Black Friday History and Statistics.” BlackFriday.com, BlackFriday.com, 10 Jan. 2019, blackfriday.com/news/black-friday-history.

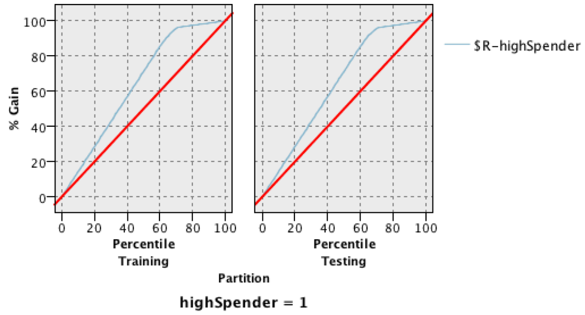
Appendix A:

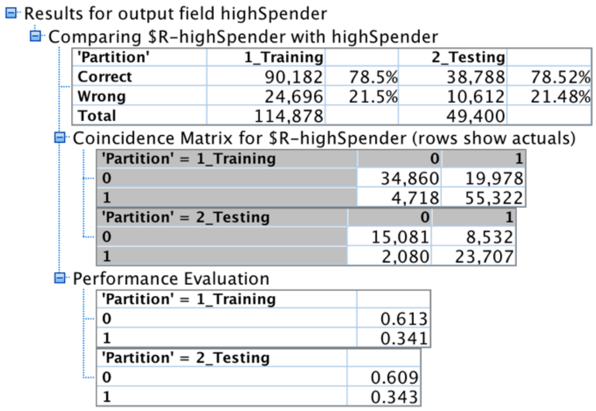
Cart Decision Tree:





Examples of Some Rules: The root is Product_Category_1. It seems that the model narrows between different product categories for product 1 and 3 to gain better results. The model does not result in a pure tree.





⊟ Results for output field highSpender

⊟ Comparing $R-highSpender with highSpender

'Partition'	1_Training		2_Testing	
Correct	90,182	78.5%	38,788	78.52%
Wrong	24,696	21.5%	10,612	21.48%
Total	114,878		49,400	

⊟ Coincidence Matrix for $R-highSpender (rows show actuals)

'Partition' = 1_Training	0	1
0	34,860	19,978
1	4,718	55,322

'Partition' = 2_Testing	0	1
0	15,081	8,532
1	2,080	23,707

⊟ Performance Evaluation

'Partition' = 1_Training	
0	0.613
1	0.341

'Partition' = 2_Testing	
0	0.609
1	0.343

Antecedent	**Consequence**	**Support**	**Confidence**
If (Product_Category_1)==3,4,5,8,11,12,13	Then HighSpender=0	21995/80325	21995/23564
If (Product_Category_1)==1,2,6,10,15^Product_Category_1=1,6,10,15	Then HighSpender=1	36333/80325	36333/48945
If (Product_Category_1)==1,2,6,10,15^ Product_Category_1=2^(Product_Category_3)==4,5,9,12,14,16	Then HighSpender=0	2354/80325	2354/4100
If (Product_Category_1)==1,2,6,10,15^ Product_Category_1=2^(Product_Category_3)==8,10,15,18	Then HighSpender=1	2324/80325	2324/3716

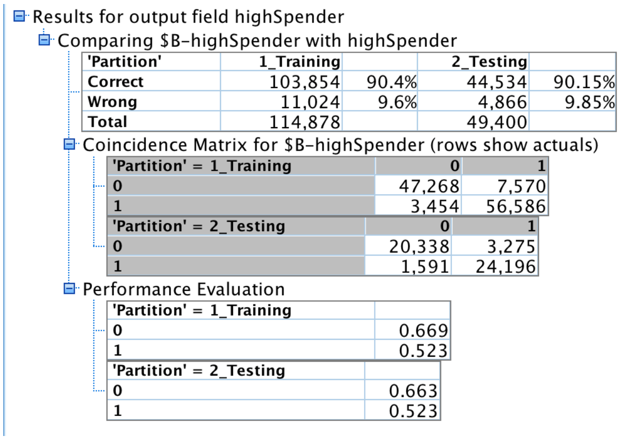
Gain Chart: At about 70% of the sample, we have around 95% gain for training and testing.

	Training:	Testing:
Accuracy: TP+TN)/(TP+TN+FP+FN)	(55322+34860)/(55322+34860+4718+19978)=78.5%	(23707+15081)/(23707+15081+2080+8532)=78.5%
Specificity: FP/(TN+FP)	19978/(34860+19978)=36.4%	(8532/(15081+8532)=36.1%
Recall: TP/(TP+FN)	55322/(55322+14718)=92.1%	23707/(23707+2080)=91.9%
Precision: TP/(TP+FP)	55322/(55322+19978)=73.5%	23707/(23707+8352)=73.5%

Appendix B

Bayesian Network

	Training:	Testing:
Accuracy: TP+TN)/(TP+TN+FP+FN)	(56586+47268)/(56586+47268+3454+7570)=90.4%	(24196+20338)/(24196+20338+3275+1591)=90.1%
Specificity: FP/(TN+FP)	7570/(7570+47268)=13.8%	3275/(3275+20338)=13.9%
Recall: TP/(TP+FN)	56586/(56586+3454)=94.2%	24196/(24196+1591)=93.8%
Precision: TP/(TP+FP)	56586/(56586+7570)=88.2%	24196/(3275+24196)=88.1%



Results for output field highSpender

Comparing $B-highSpender with highSpender

'Partition'	1_Training		2_Testing	
Correct	103,854	90.4%	44,534	90.15%
Wrong	11,024	9.6%	4,866	9.85%
Total	114,878		49,400	

Coincidence Matrix for $B-highSpender (rows show actuals)

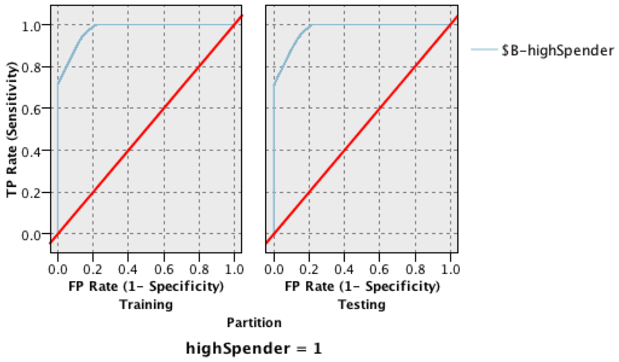
'Partition' = 1_Training	0	1
0	47,268	7,570
1	3,454	56,586

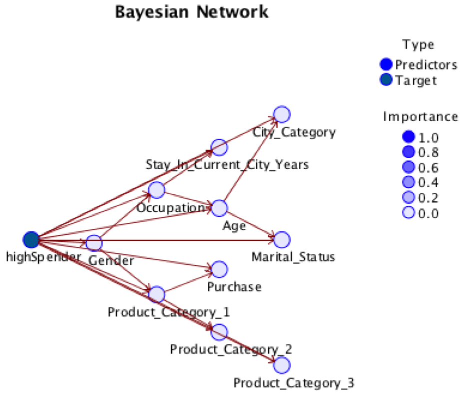
'Partition' = 2_Testing	0	1
0	20,338	3,275
1	1,591	24,196

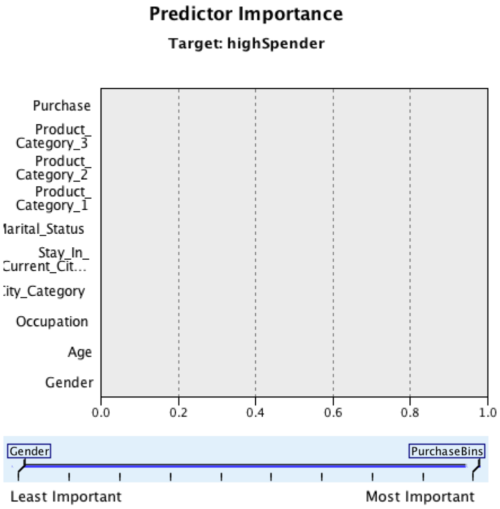
Performance Evaluation

'Partition' = 1_Training	
0	0.669
1	0.523

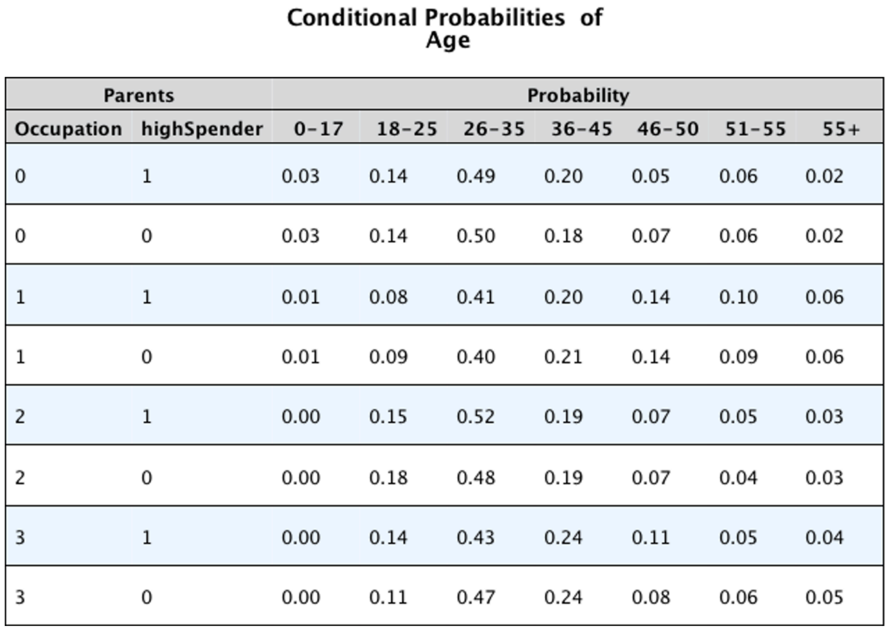
'Partition' = 2_Testing	
0	0.663
1	0.523







Example of some conditional probabilities of Age

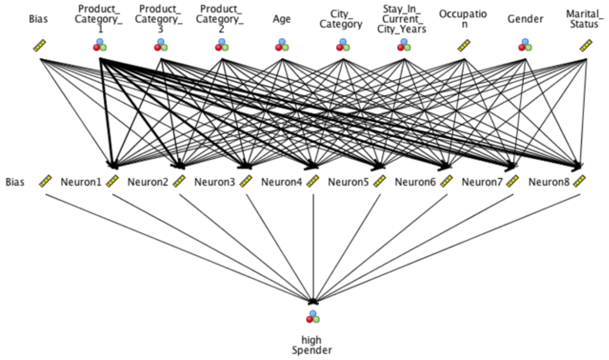


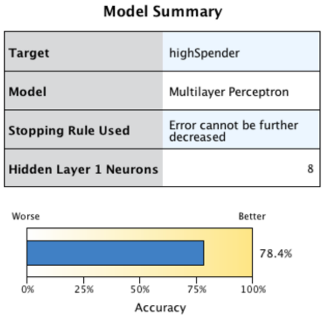
Conditional Probabilities of Age

Parents		Probability						
Occupation	highSpender	0-17	18-25	26-35	36-45	46-50	51-55	55+
0	1	0.03	0.14	0.49	0.20	0.05	0.06	0.02
0	0	0.03	0.14	0.50	0.18	0.07	0.06	0.02
1	1	0.01	0.08	0.41	0.20	0.14	0.10	0.06
1	0	0.01	0.09	0.40	0.21	0.14	0.09	0.06
2	1	0.00	0.15	0.52	0.19	0.07	0.05	0.03
2	0	0.00	0.18	0.48	0.19	0.07	0.04	0.03
3	1	0.00	0.14	0.43	0.24	0.11	0.05	0.04
3	0	0.00	0.11	0.47	0.24	0.08	0.06	0.05

Appendix C:

Neural Network:

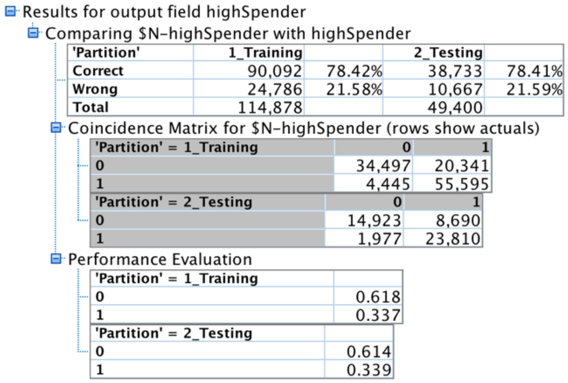




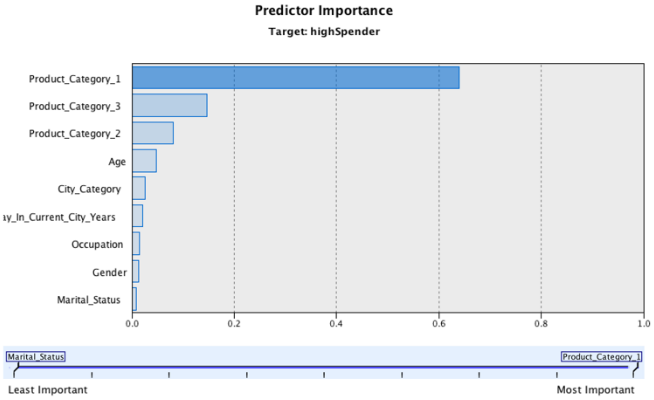
Model Summary

Target	highSpender
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	8

	Training:	Testing:
Accuracy: TP+TN)/(TP+TN+FP+FN)	(55595+34497)/(55595+34497+4445+20341)= 78.4%	(23810+14923)/(23810+14923+8690+1977)=78.4%
Specificity: FP/(TN+FP)	20341/(20341+34497)=37.1%	8690/(8690+14923)=36.8%
Recall: TP/(TP+FN)	55595/(55595+4445)=92.6%	23810/(23810+1977)=92.3%
Precision: TP/(TP+FP)	55595/(55595+20341)=73.2%	23810/(23810+8690)=73.3%



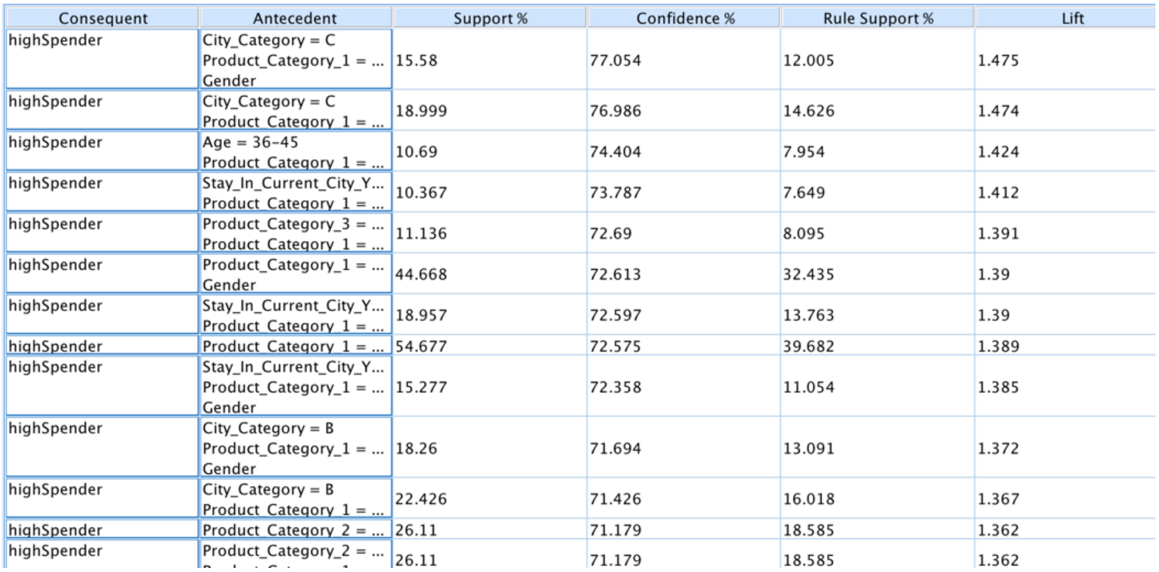
```
Results for output field highSpender
  Comparing $N-highSpender with highSpender
   'Partition'        1_Training            2_Testing
   Correct         90,092   78.42%      38,733   78.41%
   Wrong           24,786   21.58%      10,667   21.59%
   Total          114,878               49,400
  Coincidence Matrix for $N-highSpender (rows show actuals)
     'Partition' = 1_Training          0        1
     0                            34,497   20,341
     1                             4,445   55,595
     'Partition' = 2_Testing           0        1
     0                            14,923    8,690
     1                             1,977   23,810
  Performance Evaluation
     'Partition' = 1_Training
     0                             0.618
     1                             0.337
     'Partition' = 2_Testing
     0                             0.614
     1                             0.339
```



Appendix D:

Apriori Network:

Top rules

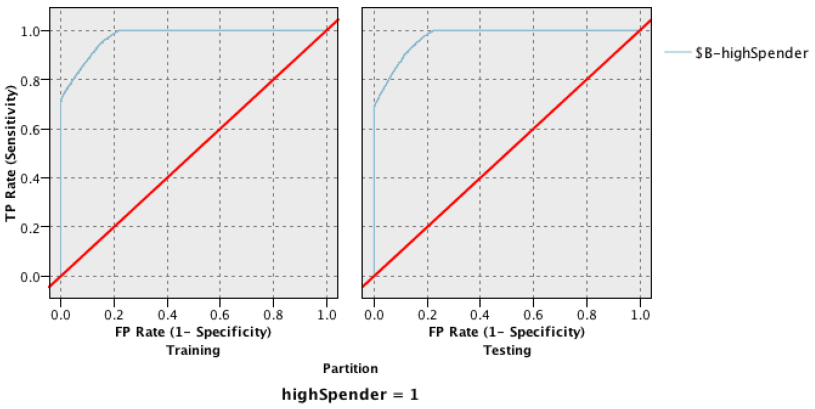


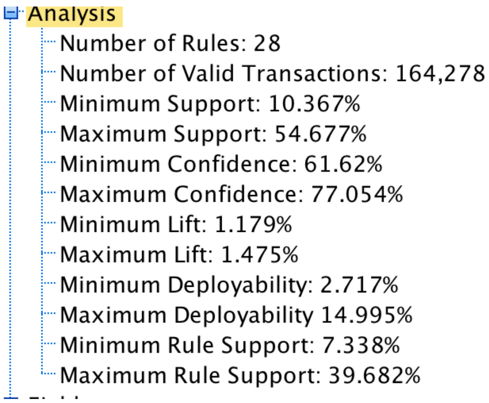
Consequent	Antecedent	Support %	Confidence %	Rule Support %	Lift
highSpender	City_Category = C Product_Category_1 = ... Gender	15.58	77.054	12.005	1.475
highSpender	City_Category = C Product_Category_1 = ...	18.999	76.986	14.626	1.474
highSpender	Age = 36-45 Product_Category_1 = ...	10.69	74.404	7.954	1.424
highSpender	Stay_In_Current_City_Y... Product_Category_1 = ...	10.367	73.787	7.649	1.412
highSpender	Product_Category_3 = ... Product_Category_1 = ...	11.136	72.69	8.095	1.391
highSpender	Product_Category_1 = ... Gender	44.668	72.613	32.435	1.39
highSpender	Stay_In_Current_City_Y... Product_Category_1 = ...	18.957	72.597	13.763	1.39
highSpender	Product_Category_1 = ...	54.677	72.575	39.682	1.389
highSpender	Stay_In_Current_City_Y... Product_Category_1 = ... Gender	15.277	72.358	11.054	1.385
highSpender	City_Category = B Product_Category_1 = ... Gender	18.26	71.694	13.091	1.372
highSpender	City_Category = B Product_Category_1 = ...	22.426	71.426	16.018	1.367
highSpender	Product_Category_2 = ...	26.11	71.179	18.585	1.362
highSpender	Product_Category_2 = ...	26.11	71.179	18.585	1.362

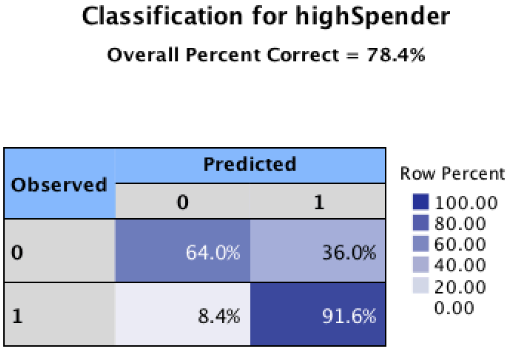
Appendix E:

Millennial Bayesian Network

	Training:	Testing:
Accuracy: (TP+TN)/(TP+TN+FP+FN)	(8359+6976)/(8359+6976+1141+583) = 89.9%	(3594+3103)/(3594+3101+524+261) = 89.5%
Specificity: FP/(TN+FP)	1141/(6976+8359) = 7.4%	524/(3103+524) = 14.4%
Recall: TP/(TP+FN)	8359/(8359+583) = 93.5%	3594/(3594+261) = 93.2%
Precision: TP/(TP+FP)	8359/(8359+1141) = 88.0%	3594/(3594+524) = 87.3%



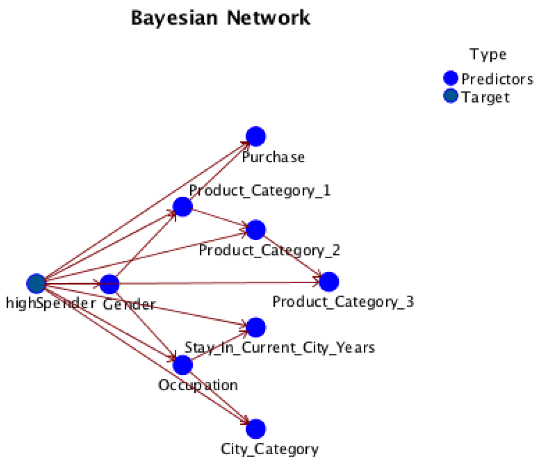

Analysis
- Number of Rules: 28
- Number of Valid Transactions: 164,278
- Minimum Support: 10.367%
- Maximum Support: 54.677%
- Minimum Confidence: 61.62%
- Maximum Confidence: 77.054%
- Minimum Lift: 1.179%
- Maximum Lift: 1.475%
- Minimum Deployability: 2.717%
- Maximum Deployability 14.995%
- Minimum Rule Support: 7.338%
- Maximum Rule Support: 39.682%

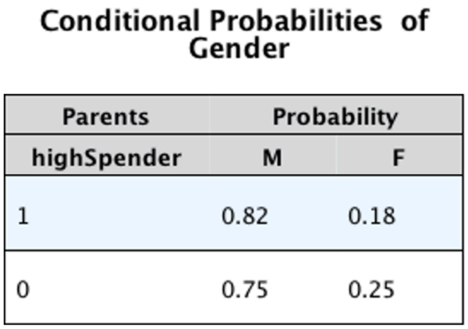

Classification for highSpender

Overall Percent Correct = 78.4%

Observed	Predicted 0	Predicted 1
0	64.0%	36.0%
1	8.4%	91.6%

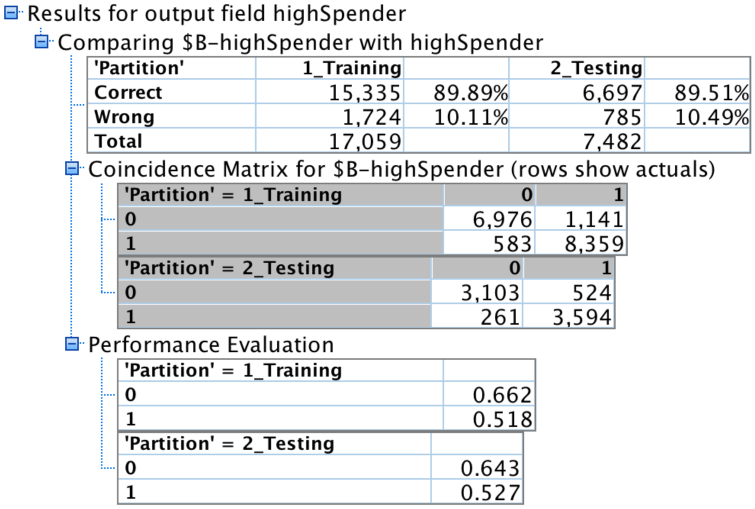
Row Percent
- 100.00
- 80.00
- 60.00
- 40.00
- 20.00
- 0.00





Conditional Probabilities of Gender

Parents	Probability	
highSpender	M	F
1	0.82	0.18
0	0.75	0.25



Results for output field highSpender

Comparing $B-highSpender with highSpender

'Partition'	1_Training		2_Testing	
Correct	15,335	89.89%	6,697	89.51%
Wrong	1,724	10.11%	785	10.49%
Total	17,059		7,482	

Coincidence Matrix for $B-highSpender (rows show actuals)

'Partition' = 1_Training	0	1
0	6,976	1,141
1	583	8,359

'Partition' = 2_Testing	0	1
0	3,103	524
1	261	3,594

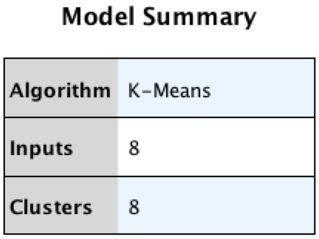
Performance Evaluation

'Partition' = 1_Training	
0	0.662
1	0.518

'Partition' = 2_Testing	
0	0.643
1	0.527

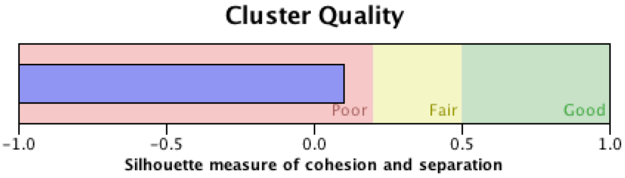
Appendix F:

Millennial K Means

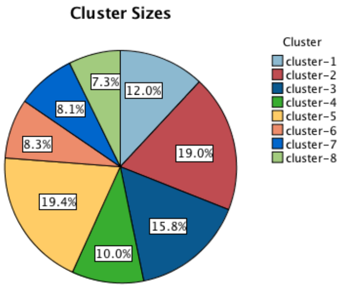


Model Summary

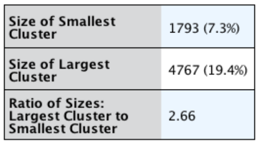
Algorithm	K-Means
Inputs	8
Clusters	8



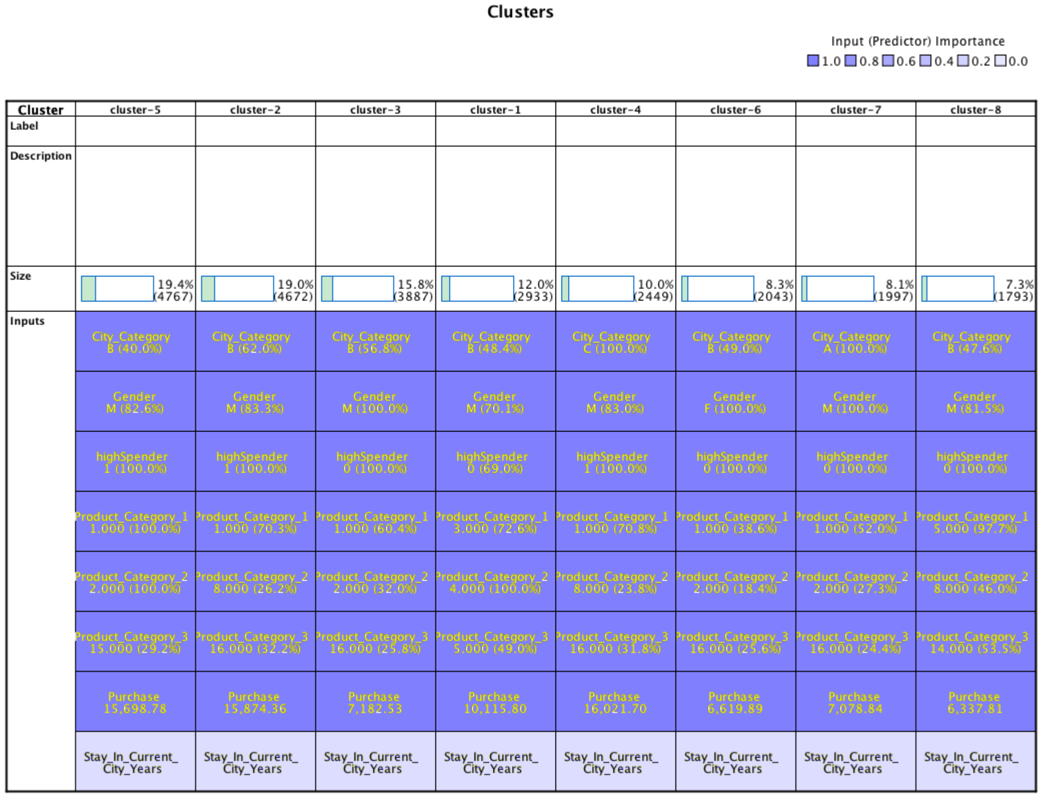
Cluster Quality



Cluster Sizes



Size of Smallest Cluster	1793 (7.3%)
Size of Largest Cluster	4767 (19.4%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	2.66



Clusters

Input (Predictor) Importance
■1.0 ■0.8 ■0.6 ■0.4 □0.2 □0.0

Cluster	cluster-5	cluster-2	cluster-3	cluster-1	cluster-4	cluster-6	cluster-7	cluster-8
Label								
Description								
Size	19.4% (4767)	19.0% (4672)	15.8% (3887)	12.0% (2933)	10.0% (2449)	8.3% (2043)	8.1% (1997)	7.3% (1793)
Inputs	City_Category B (40.0%)	City_Category B (62.0%)	City_Category B (56.8%)	City_Category B (48.4%)	City_Category C (100.0%)	City_Category B (49.0%)	City_Category A (100.0%)	City_Category B (47.6%)
	Gender M (82.6%)	Gender M (83.3%)	Gender M (100.0%)	Gender M (70.1%)	Gender M (83.0%)	Gender F (100.0%)	Gender M (100.0%)	Gender M (81.5%)
	highSpender 1 (100.0%)	highSpender 1 (100.0%)	highSpender 0 (100.0%)	highSpender 0 (69.0%)	highSpender 1 (100.0%)	highSpender 0 (100.0%)	highSpender 0 (100.0%)	highSpender 0 (100.0%)
	Product_Category_1 1.000 (100.0%)	Product_Category_1 1.000 (70.3%)	Product_Category_1 1.000 (60.4%)	Product_Category_1 3.000 (72.6%)	Product_Category_1 1.000 (70.8%)	Product_Category_1 1.000 (38.6%)	Product_Category_1 1.000 (52.0%)	Product_Category_1 5.000 (97.7%)
	Product_Category_2 2.000 (100.0%)	Product_Category_2 8.000 (26.2%)	Product_Category_2 2.000 (32.0%)	Product_Category_2 4.000 (100.0%)	Product_Category_2 8.000 (23.8%)	Product_Category_2 2.000 (18.4%)	Product_Category_2 2.000 (27.3%)	Product_Category_2 8.000 (46.0%)
	Product_Category_3 15.000 (29.2%)	Product_Category_3 16.000 (32.2%)	Product_Category_3 16.000 (25.8%)	Product_Category_3 5.000 (49.0%)	Product_Category_3 16.000 (31.8%)	Product_Category_3 16.000 (25.6%)	Product_Category_3 16.000 (24.4%)	Product_Category_3 14.000 (53.5%)
	Purchase 15,698.78	Purchase 15,874.36	Purchase 7,182.53	Purchase 10,115.80	Purchase 16,021.70	Purchase 6,619.89	Purchase 7,078.84	Purchase 6,337.81
	Stay_In_Current_City_Years	Stay_In_Current_City_Years	Stay_In_Current_City_Years	Stay_In_Current_City_Years	Stay_In_Current_City_Years	Stay_In_Current_City_Years	Stay_In_Current_City_Years	Stay_In_Current_City_Years