ILLINOIS INSTITUTE OF TECHNOLOGY

ITMD 525 – DATA MINING FINAL PROJECT

Credit Card Transaction Categorization using Text Analytics

Team Members:

Anusha Bondili – A20356204

Manikandan Ganesh – A20355226

Jay Thakker – A20359548

ILLINOIS INSTITUTE OF TECHNOLOGY

TABLE OF CONTENTS

1. INTRODUCTION

- Text Analytics a.k.a. Text Mining refers to the process of deriving high-quality information from text.
- Text Mining usually involves parsing of input text data to find out meaningful patterns and interpreting those patterns as a form of an output.
- Areas where text mining is implemented is Text Categorization, Sentiment Analysis, Text Clustering, Document Summarizing and many more.
- In this project, we have used text mining for Categorization of Credit Card Transactions.

2. OBJECTIVES & APPLICATIONS

- **Objectives**
  - Carry out text mining on transaction description to identify word patterns belonging to different categories
  - Using the text patterns, bucket the credit card transaction data in 38 different categories.
- **Applications**
  - A user would be able to view their spending in an understandable format.
  - A bank would be able to analyze the expenses in every category and come up with different credit card offers or cash back rewards system.

ILLINOIS INSTITUTE OF TECHNOLOGY

3. SNAPSHOT OF THE DATASET

- Creditdata_Category

DESCRIPTION	LEDGER_ENTRY	PROPOSED_CATEGORY
DIVERSIFIED VENDING LLC WALLINGFORD CT	debit	Groceries
NATIONAL CAR TOLLS	debit	Travel
FORSBERG FINE WINE & S CHARLESTON SC XXXXX USA	debit	Groceries
Interest on Purchases	debit	Service Charges/Fees
OFFUTT NEW MAIN STORE	debit	Personal/Family

- Train Dataset
- Observations: 501,449.
- Categories assigned manually.

- CreditCard_Actual

DESCRIPTION	TRANSACTION_DATE	AMOUNT	TYPE	Category
STOP & SHOP #569 OYSTER BAY NY	7/29/2014	86.3	debit	
NTTA CUST SVC ONLINE PLANO TX	1/30/2014	34.43	debit	
NETFLIX.COM XXXXXXXXXX CA	2/12/2015	7.99	debit	
PRESIDENT OFFICE 00-08 BROOKLYN	11/10/2014	20	debit	
WHOLEFDS RMD XXXXX 0REDMOND XXXXXXXXXXX	1/2/2015	98.42	debit	

- Test Dataset
- Observations: 500,001.
- Using the train dataset, assign the categories.

4. CHALLENGES & APPROACH

- **Challenges**
  - Tackle parts of speech present in transaction description to identify patterns
  - Inconsistent data having spelling mistakes
  - We have "Other Expenses" as a category which contains transactions that do not fall under any of the major categories.
  - Identical terms falling under different categories

ILLINOIS INSTITUTE OF TECHNOLOGY

4. CHALLENGES & APPROACH

- **Approach**
  - We extract only NOUNS from every transaction description and eliminate different parts of speech, locations, English stop words, zip codes and abbreviations.
  - Miscellaneous nouns present in the same transaction description of the training data set are associated with the category to form different rules.
  - The transaction description of the test dataset is parsed against these rules to predict the category of the transactions present in the test data.
  - To associate a transaction to a specific category, probability of the associated rule to each category is calculated and the category with the highest probability value is predicted as the associated transaction.

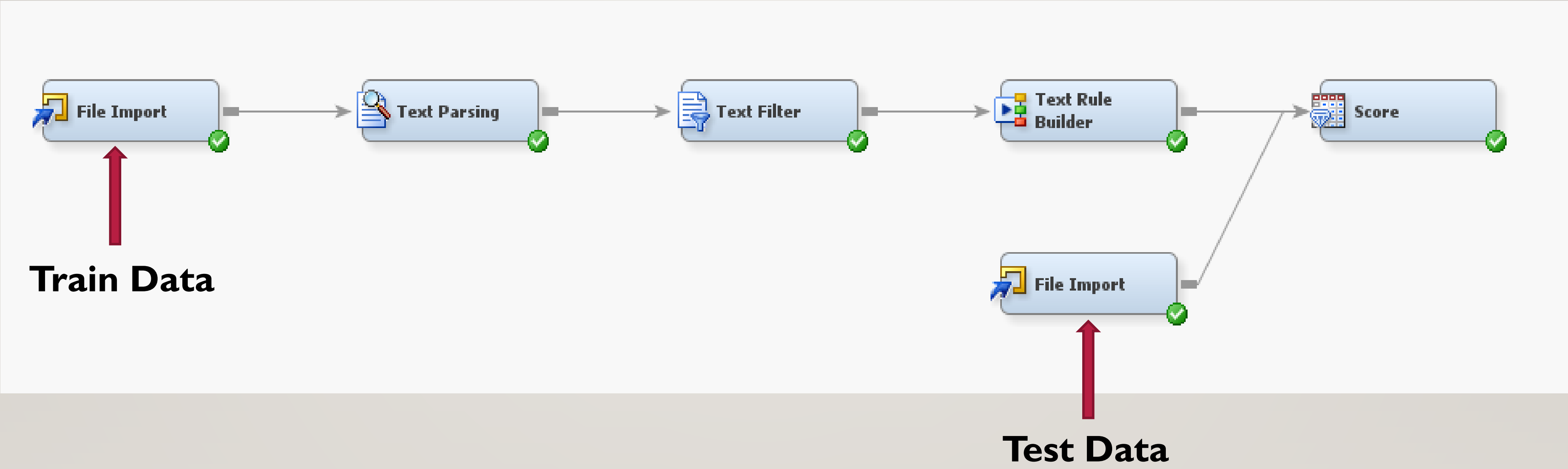
5. IMPLEMENTATION

- Text Mining module of SAS Enterprise Miner is used to create a transformation to achieve this objective.
- Different nodes of this module used in the system are:
  - File Import – Import the file (.xlsx/.sas7bdat)
  - Text Parsing – Extract nouns and eliminate other parts of speech
  - Text Filter – Weigh the terms using Mutual Information and TF.IDF
  - Text Rule Builder – Build rules to identify different text patterns
  - Score Node – Generates a score for transaction w.r.t. every category

ILLINOIS INSTITUTE OF TECHNOLOGY

5. IMPLEMENTATION

- The transformation created using these components is below:



5. IMPLEMENTATION

- To achieve highest accuracy, we split the train data in 2 parts (80% train – 20% test).
- Eliminate all transactions that fall under "Other Expenses" category.
- Document Term Matrix is formed by assigning Mutual Information weights for each term associated with documents.

Terms

Term	Role	Attribute	Status	Weight	Imported Frequency	Freq	Number of Imported Documents	# Docs	Rank	Parent/Child Status	Parent ID
com	Noun	Alpha	Drop	0.000	25199	25199	21432	21432	1		35917
+ amazon	Noun	Alpha	Keep	0.738	11153	11153	11079	11079	2+		23026
+ store	Noun	Alpha	Keep	0.880	9578	9578	9276	9276	3+		7306
s	Noun	Alpha	Drop	0.000	9348	9348	9260	9260	4		35893
interest	Noun	Alpha	Keep	1.171	8283	8283	8081	8081	5		15688
wal-mart	Noun	Mixed	Keep	0.814	5501	5501	5501	5501	6		21622
+ bill	Noun	Alpha	Keep	0.861	5128	5128	5119	5119	7+		10900
oil	Noun	Alpha	Keep	1.083	4896	4896	4896	4896	8		11980
target	Noun	Alpha	Keep	0.814	4615	4615	4615	4615	9		18856
shell	Noun	Alpha	Keep	1.083	4408	4408	4407	4407	10		11891
+ cafe	Noun	Alpha	Keep	0.638	4341	4341	4335	4335	11+		26690
shell oil	Noun Group	Alpha	Keep	1.083	4218	4218	4218	4218	12		30549
+ fee	Noun	Alpha	Keep	1.587	4032	4032	3993	3993	13+		9239
+ restaurant	Noun	Alpha	Keep	0.644	3974	3974	3939	3939	14+		1879
+ apple	Noun	Alpha	Keep	1.095	3671	3671	3646	3646	15+		3456
depot	Noun	Alpha	Keep	1.429	3644	3644	3644	3644	16		30576
+ pharmacy	Noun	Alpha	Keep	1.404	3613	3613	3589	3589	17+		12668
+ market	Noun	Alpha	Keep	0.888	3274	3274	3269	3269	18+		10840
apple itunes store xxx...	Noun Group	Mixed	Keep	1.149	2670	2670	2670	2670	19		28500
amazon.com	Noun	Mixed	Keep	0.814	2663	2663	2663	2663	20		17270
+ payment	Noun	Alpha	Keep	1.214	2348	2348	2348	2348	21+		4730
+ food	Noun	Alpha	Keep	0.724	2320	2320	2316	2316	22+		23109
+ fuel	Noun	Alpha	Keep	0.888	2261	2261	2236	2236	23+		16001
+ pizza	Noun	Alpha	Keep	0.645	2221	2221	2219	2219	24+		29294
+ subway	Noun	Alpha	Keep	0.647	2193	2193	2192	2192	25+		1314
+ pa	Noun	Alpha	Keep	0.711	2184	2184	2183	2183	26+		18955
+ city	Noun	Alpha	Keep	0.602	2168	2168	2148	2148	27+		4857
+ service	Noun	Alpha	Keep	1.126	2132	2132	2005	2005	28+		21193
foreign tran chg	Noun Group	Alpha	Keep	1.171	2002	2002	2002	2002	29		5067
chevron	Noun	Alpha	Keep	1.083	1984	1984	1976	1976	30		1316
+ park	Noun	Alpha	Keep	0.685	2037	2037	1967	1967	31+		29010
+ shop	Noun	Alpha	Keep	1.146	1970	1970	1965	1965	32+		12033
+ trader	Noun	Alpha	Keep	0.922	1959	1959	1959	1959	33+		7781
joe	Noun	Alpha	Keep	0.917	1956	1956	1955	1955	34		21572
+ trader joe	Noun Group	Alpha	Keep	0.928	1932	1932	1932	1932	35+		8891
supercenter	Noun	Alpha	Keep	0.814	1915	1915	1915	1915	36		30058
+ charge	Noun	Alpha	Keep	1.169	2044	2044	1842	1842	37+		29353
+ ma	Noun	Alpha	Keep	1.000	1727	1727	1726	1726	38+		5807
fort	Noun	Alpha	Keep	0.747	1756	1756	1688	1688	39		7650
c	Noun	Alpha	Drop	0.000	1684	1684	1623	1623	40		35890
grill	Noun	Alpha	Keep	0.646	1602	1602	1601	1601	41		8006
+ burger	Noun	Alpha	Keep	0.647	1567	1567	1565	1565	42+		5024
+ hill	Noun	Alpha	Keep	1.152	1571	1571	1541	1541	43+		20074
+ taco	Noun	Alpha	Keep	0.647	1527	1527	1524	1524	44+		19126
+ lake	Noun	Alpha	Keep	0.438	1487	1487	1445	1445	45+		5478
membership	Noun	Alpha	Keep	2.076	1426	1426	1423	1423	46		10526
chipotle	Noun	Alpha	Keep	0.648	1422	1422	1422	1422	47		24579
+ king	Noun	Alpha	Keep	0.472	1365	1365	1361	1361	48+		20401
+ beach	Noun	Alpha	Keep	0.362	1317	1317	1285	1285	49+		8251
transaction	Noun	Alpha	Keep	1.162	1270	1270	1268	1268	50		22226
taxi	Noun	Alpha	Keep	1.210	1238	1238	1235	1235	51		21033
+ island	Noun	Alpha	Keep	0.957	1234	1234	1230	1230	52+		11100
+ la	Noun	Alpha	Keep	0.301	1202	1202	1178	1178	53+		22076
+ car	Noun	Alpha	Keep	0.879	1182	1182	1170	1170	54+		16596

5. IMPLEMENTATION

Rules Obtained

Target Value	Rule #	Rule	Precision	Recall	F1 score	True Positive/Total
POSTAGE/SHIPPING	1258	horseshoe	[illegible]	[illegible]	[illegible]	[illegible]
POSTAGE/SHIPPING	1259	new hyde parkny	80.38%	6.83%	12.58%	2/17
POSTAGE/SHIPPING	1260	north myrtle sc	79.72%	6.87%	12.64%	1/8
POSTAGE/SHIPPING	1261	new brunswicknj	78.34%	6.91%	12.70%	1/8
POSTAGE/SHIPPING	1262	hero	77.73%	6.95%	12.76%	1/8
POSTAGE/SHIPPING	1263	fond	76.79%	6.99%	12.81%	1/8
POSTAGE/SHIPPING	1264	satellite	74.68%	7.07%	12.92%	2/19
POSTAGE/SHIPPING	1265	office & ~depot	74.68%	7.07%	12.92%	52/408
RENT	1266	u-haul	100.0%	46.59%	63.56%	198/198
RENT	1267	extra space stora	100.0%	55.06%	71.02%	36/36
RENT	1268	rental & purchase	100.0%	60.71%	75.55%	24/24
RENT	1269	(800)528-0463	100.0%	63.29%	77.52%	11/11
RENT	1270	extra space stor	100.0%	65.65%	79.26%	10/10
RENT	1271	extra space storage	100.0%	67.06%	80.28%	6/6
RENT	1272	household	99.31%	67.53%	80.39%	2/4
RENT	1273	mini	99.31%	67.53%	80.39%	4/24

6. ANALYSIS

- Accuracy is determined using the following parameters:
  - Generalization Error – Determines the predicted probability for rules on untrained data

$$I[f_n] = \int_{X\times Y} V(f_n(x), y)\rho(x, y)dxdy,$$

$$I_S[f_n] = \frac{1}{n}\sum_{i=1}^{n} V(f_n(x_i), y_i)$$

$$G = I[f_n] - I_S[f_n]$$

  - Purity of rules – Higher values result in fewer, purer rules. Lower values cover more instances
  - Exhaustiveness – Controls the number of terms formed as a conjunction with the existing rule using AND (&) operator

ILLINOIS INSTITUTE OF TECHNOLOGY

6. ANALYSIS

- **Configuration #1**

Generalization Error	Purity of Rules	Exhaustiveness
Medium	Medium	Low

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
PROPOSED C...	PROPOSED C...	ASE	Average Squared Error	0.01338	.	.
PROPOSED C...	PROPOSED C...	DIV	Divisor for ASE	7657325	.	.
PROPOSED C...	PROPOSED C...	MAX	Maximum Absolute Error	0.964784	.	.
PROPOSED C...	PROPOSED C...	NOBS	Sum of Frequencies	306293	.	.
PROPOSED C...	PROPOSED C...	RASE	Root Average Squared Error	0.115674	.	.
PROPOSED C...	PROPOSED C...	SSE	Sum of Squared Errors	102458.2	.	.
PROPOSED C...	PROPOSED C...	DISF	Frequency of Classified Cases	306293	.	.
PROPOSED C...	PROPOSED C...	MISC	Misclassification Rate	0.374181	.	.
PROPOSED C...	PROPOSED C...	WRONG	Number of Wrong Classifications	114609	.	.

ILLINOIS INSTITUTE OF TECHNOLOGY

6. ANALYSIS

- **Configuration #2**

Generalization Error	Purity of Rules	Exhaustiveness
Very Low	Low	Low

Target	Target Label	Fit Statistics	Statistics Label	Validation	Test	Train
PROPOSED CAT...	PROPOSED CAT...	ASE	Average Squared Error	.	.	0.013206
PROPOSED CAT...	PROPOSED CAT...	DIV	Divisor for ASE	.	.	7657325
PROPOSED CAT...	PROPOSED CAT...	MAX	Maximum Absolute Error	.	.	0.971622
PROPOSED CAT...	PROPOSED CAT...	NOBS	Sum of Frequencies	.	.	306293
PROPOSED CAT...	PROPOSED CAT...	RASE	Root Average Squared Error	.	.	0.114916
PROPOSED CAT...	PROPOSED CAT...	SSE	Sum of Squared Errors	.	.	101120.7
PROPOSED CAT...	PROPOSED CAT...	DISF	Frequency of Classified Cases	.	.	306293
PROPOSED CAT...	PROPOSED CAT...	MISC	Misclassification Rate	.	.	0.373858
PROPOSED CAT...	PROPOSED CAT...	WRONG	Number of Wrong Classifications	.	.	114510

6. ANALYSIS

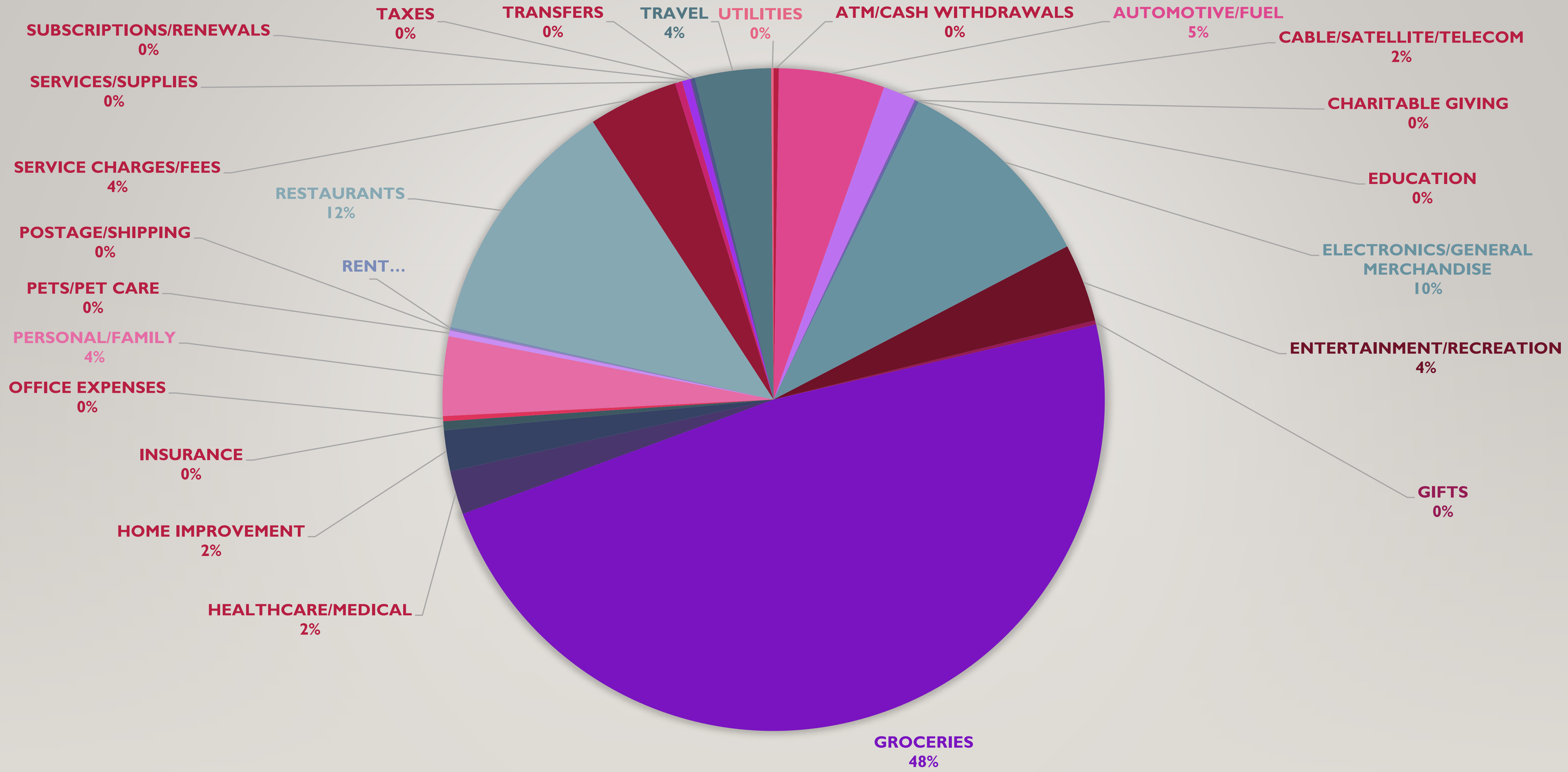
- **Configuration #3**

Generalization Error	Purity of Rules	Exhaustiveness
Very Low	Medium	Low

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
PROPOSED CA...	PROPOSED CA...	ASE	Average Squared ...	0.013284	.	.
PROPOSED CA...	PROPOSED CA...	DIV	Divisor for ASE	7657325	.	.
PROPOSED CA...	PROPOSED CA...	MAX	Maximum Absolut...	0.970906	.	.
PROPOSED CA...	PROPOSED CA...	NOBS	Sum of Frequencies	306293	.	.
PROPOSED CA...	PROPOSED CA...	RASE	Root Average Squ...	0.115257	.	.
PROPOSED CA...	PROPOSED CA...	SSE	Sum of Squared E...	101722	.	.
PROPOSED CA...	PROPOSED CA...	DISF	Frequency of Clas...	306293	.	.
PROPOSED CA...	PROPOSED CA...	MISC	Misclassification ...	0.37406	.	.
PROPOSED CA...	PROPOSED CA...	WRONG	Number of Wrong ...	114572	.	.

DEMO

ILLINOIS INSTITUTE OF TECHNOLOGY


TAXES
0%
TRANSFERS
0%
TRAVEL
4%
UTILITIES
0%
ATM/CASH WITHDRAWALS
0%
AUTOMOTIVE/FUEL
5%
SUBSCRIPTIONS/RENEWALS
0%
CABLE/SATELLITE/TELECOM
2%
SERVICES/SUPPLIES
0%
CHARITABLE GIVING
0%
SERVICE CHARGES/FEES
4%
EDUCATION
0%
RESTAURANTS
12%
POSTAGE/SHIPPING
0%
ELECTRONICS/GENERAL MERCHANDISE
10%
RENT...
PETS/PET CARE
0%
PERSONAL/FAMILY
4%
ENTERTAINMENT/RECREATION
4%
OFFICE EXPENSES
0%
INSURANCE
0%
GIFTS
0%
HOME IMPROVEMENT
2%
HEALTHCARE/MEDICAL
2%
GROCERIES
48%

QUESTIONS

ILLINOIS INSTITUTE OF TECHNOLOGY