MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Massey University, Multivariate Analysis for Big Data

Exploring the Customer Churning with multivariate statistics analysis

ID: 24002398

Name: Thai Minh Nguyen

Course Coordinator: Dr Matthew Pawley

Table of Contents

Abstract *3*

1. Introduction *4*

2. Data Source and Methodology *4*

2.1. Data Source 4

2.2. Methodology for analysis 5

3. Data Analysis *5*

Cluster Analysis 5

Correspondence Analysis 7

Canonical Correlation Analysis 9

Discriminant Analysis 10

Partial Least Squares Analysis 12

4. Discussion *14*

5. Conclusion *15*

6. Appendix *16*

Bibliography *21*

Abstract

The analysis examined a fictional telecom company's customer churn dataset, utilizing multivariate statistical techniques including Cluster Analysis, Correspondence Analysis, Canonical Correlation Analysis, Discriminant Analysis, and Partial Least Squares analysis. Results revealed a prominent customer group comprising young individuals opting for Fiber Optic internet service with month-to-month contracts, exhibiting infrequent need for tech support. Younger customers with partners but no dependents were found to be more likely to use streaming services, while those with dependents were less likely to use internet services. Discriminant analysis, utilizing all variables except the insignificant ones (PhoneService and OnlineBackup), resulted in a misclassification rate of 24.22%. Key predictors of churn included relationship status (having partners and dependents or not), tenure, streaming TV and movie usage, paperless billing, monthly charges, fiber optic service adoption, month-to-month or one-year contracts, and electronic check payment method. OnlineSecurity, DeviceProtection, and StreamingTV exhibited a strong positive correlation with tenure.

1. Introduction

The main objective of this project was to gain a better understanding of the relationship between customers' attributes and the services they use, specifically regarding their likelihood to churn, using multivariate statistical techniques. The techniques utilized in this report include Cluster Analysis, Correspondence Analysis, Canonical Correlation Analysis, Discriminant Analysis, and Partial least squares analysis.

2. Data Source and Methodology

2.1. Data Source

The dataset utilized for analysis tracks a fictional telco company's customer churn based on various factors. It is comprised of 21 columns and 7032 rows. Each row represents a customer, each column contains the customer's attributes described in the column (Kaggle, n.d.).

Variables	Definition
customerID	Customer ID
gender	Whether the customer is a male or a female
SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)
Partner	Whether the customer has a partner or not (Yes, No)
Dependents	Whether the customer has dependents or not (Yes, No)
tenure	Number of months the customer has stayed with the company
PhoneService	Whether the customer has a phone service or not (Yes, No)
MultipleLines	Whether the customer has multiple lines or not (Yes, No, No phone service)
InternetService	Customer's internet service provider (DSL, Fiber optic, No)
OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)
OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)
DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)
TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)
StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)
StreamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)
Contract	The contract term of the customer (Month-to-month, One year, Two year)
PaperlessBilling	Whether the customer has paperless billing or not (Yes, No)
PaymentMethod	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card

MonthlyCharges	The amount charged to the customer monthly
TotalCharges	The total amount charged to the customer check
Churn	Whether the customer churned or not (Yes or No)

The data was downloaded as a CSV file and imported to the SAS dataset. Moreover, because most of the columns were categories, the other version of the dataset was generated by using Python. The other version converted the column with Yes or No to 0 and 1. If there were more than 2 options, I converted these columns to dummy variables. For example, The Contract column has 3 unique values Month-to-month, One year, and Two year, which were converted to Contract_Month-to-month, Contract_OneYear, and Contract_TwoYear with values 0 and 1.

2.2. Methodology for analysis

Most of the techniques used for analysis are done by SAS procedure. It included:

- Proc cluster
- Proc corresp
- Proc cancorr
- Proc discrim
- Proc pls

3. Data Analysis

Cluster Analysis

Cluster analysis was conducted to find the group of customers that are the most similar in terms of their attributes. The cluster analysis uses the CLUSTER procedure to hierarchically cluster the observations in a SAS dataset by using the Centroid Method (SAS, The CLUSTER Procedure, n.d.). Because the data consists of interval and nominal data types, I use the DISTANCE procedure to calculate the distance matrix with the method DGOWER, which is equal to 1 minus GOWER (SAS, The DISTANCE Procedure, n.d.) before inputting the distance matrix to proc cluster in SAS.

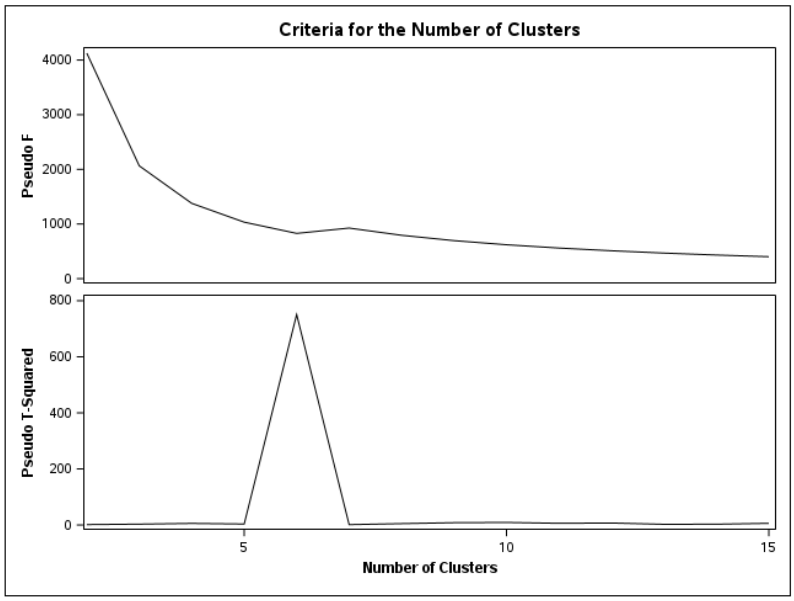

Figure 1: Criteria for the Number of Clusters

By using the pseudo F statistics (Caliński, 1974) and t-square statistics (Duda, 1974), I chose 7 clusters. From the group bar chart Figure 2 below, we can see that Cluster 1 (orange bar) is the group of young people, who choose Fiber Optic Internet service with month-to-month contracts and they do not often need Tech Support.

Figure 2: Cluster Visualization

Correspondence Analysis

The Correspondence Analysis (SAS, The CORRESP Procedure, n.d.) was used to explore the association between customer attributes such as Gender, Senior, having Dependents, and having a Partner with entertainment purposes such as streaming TV and movies. First, determine that there is a statistically significant association. The Cramer's V (IBM, 2024) of the two tables is larger than 0.7, which means there are strong associations between customer attributes and streaming activities.

From Figure 4 and Figure 5, there is a clear distinction between the points using streaming services, not using streaming, and no internet services. In the row point, the No point was close to the centroid, it contributes almost nothing to the inertia of dimension one. The Yes and No internet service points are far from the centroid. They make relatively large contributions to the chi-square statistic and the inertia of dimension one. The two interpretations of dimension one show the association between being younger, having a partner but not having dependents with using streaming services, and being younger and having dependents with no internet service. The dimension 2 shows the association between both young and senior females who having no partner with no use streaming TV and movies.

Statistics for Table of group by StreamingTV

Statistic	DF	Value	Prob
Chi-Square	16	7654.6427	<.0001
Likelihood Ratio Chi-Square	16	7851.3362	<.0001
Mantel-Haenszel Chi-Square	1	125.4933	<.0001
Phi Coefficient		1.0425	
Contingency Coefficient		0.7217	
Cramer's V		0.7372	

Sample Size = 7043

Statistics for Table of group by StreamingMovies

Statistic	DF	Value	Prob
Chi-Square	16	7688.3594	<.0001
Likelihood Ratio Chi-Square	16	7878.2491	<.0001
Mantel-Haenszel Chi-Square	1	140.6634	<.0001
Phi Coefficient		1.0448	
Contingency Coefficient		0.7224	
Cramer's V		0.7388	

Sample Size = 7043

Figure 3: Cramer's V

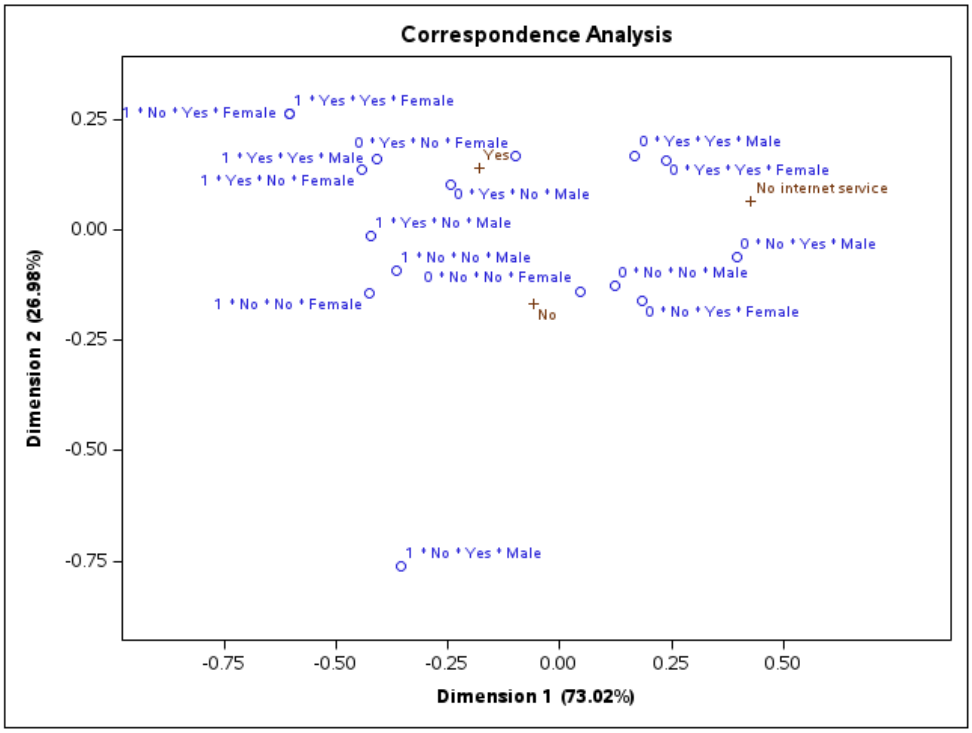

Figure 4: Correspondence Analysis with StreamingTV

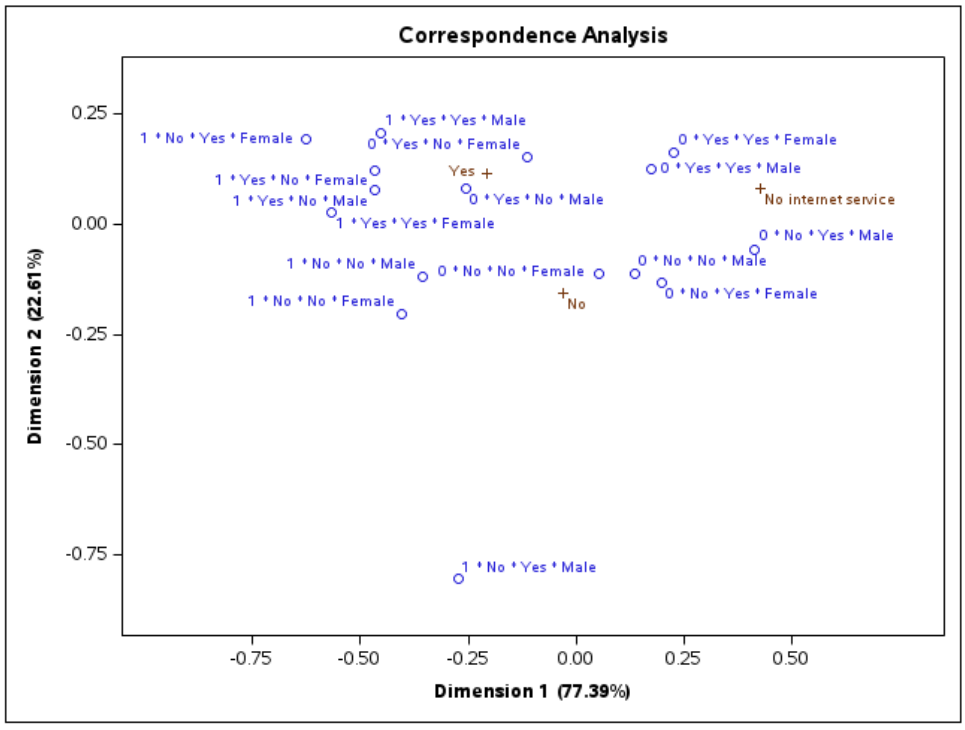

Figure 5: Correspondence Analysis with StreamingMovies

Canonical Correlation Analysis

The Canonical Correlation Analysis (SAS, The CANCORR Procedure, n.d.) was used to explore the relationship between the services, that the telecom company provided such as online security protection, online backup, streaming, internet service, and phone service with the loyalty of customers, such as the tenure and churn variables.

Multivariate Statistics and F Approximations					
S=2 M=3 N=3515					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.54946190	272.73	18	14064	<.0001
Pillai's Trace	0.50193626	261.83	18	14066	<.0001
Hotelling-Lawley Trace	0.72641971	283.76	18	11716	<.0001
Roy's Greatest Root	0.55911465	436.92	9	7033	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

Figure 6: CCA Hypothesis Test

Because the p-value is lower than 0.0001, we can reject the null hypothesis that all the canonical correlations are simultaneously zero from the multivariate statistics.

Canonical Structure

Correlations Between the SERVICE and Their Canonical Variables		
	S1	S2
PhoneService	0.0091	0.0454
OnlineSecurity	0.3481	0.4020
OnlineBackup	0.3646	0.5743
DeviceProtection	0.3616	0.6037
TechSupport	0.3441	0.4096
StreamingTV	0.2550	0.7490
StreamingMovies	0.2616	0.7508
PhoneService	0.0091	0.0454
InternetService_DSL	0.0600	-0.3256
InternetService_Fiber optic	-0.0683	0.8839

Correlations Between the LOYALTY and Their Canonical Variables		
	L1	L2
tenure	0.9838	0.1792
Churn_binary	-0.5142	0.8577

Correlations Between the SERVICE and the Canonical Variables of the LOYALTY		
	L1	L2
PhoneService	0.0055	0.0172
OnlineSecurity	0.2085	0.1522
OnlineBackup	0.2183	0.2174
DeviceProtection	0.2166	0.2285
TechSupport	0.2061	0.1551
StreamingTV	0.1527	0.2836
StreamingMovies	0.1566	0.2842
PhoneService	0.0055	0.0172
InternetService_DSL	0.0359	-0.1233
InternetService_Fiber optic	-0.0409	0.3346

Correlations Between the LOYALTY and the Canonical Variables of the SERVICE		
	S1	S2
tenure	0.5891	0.0678
Churn_binary	-0.3079	0.3247

Figure 7: Canonical Structure

Canonical structure output in Figure 7, shows that OnlineBackup, DeviceProtection, StreamingTV, Movies, and Fiber Optic Internet Services were highly correlated with S2, and Tenure and Churn had a high correlation with L1 and L2 respectively. Only the Tenure had a strong positive correlation with S1, while there is not any strong correlation between Service variables with L1 or L2.

Canonical Redundancy Analysis

Standardized Variance of the SERVICE Explained by					
	Their Own Canonical Variables			The Opposite Canonical Variables	
Canonical Variable Number	Proportion	Cumulative Proportion	Canonical R-Square	Proportion	Cumulative Proportion
1	0.0645	0.0645	0.3586	0.0231	0.0231
2	0.3040	0.3685	0.1433	0.0436	0.0667

Standardized Variance of the LOYALTY Explained by					
	Their Own Canonical Variables			The Opposite Canonical Variables	
Canonical Variable Number	Proportion	Cumulative Proportion	Canonical R-Square	Proportion	Cumulative Proportion
1	0.6162	0.6162	0.3586	0.2210	0.2210
2	0.3838	1.0000	0.1433	0.0550	0.2760

Figure 8: Redundancy Analysis

The standardized variance output of the Redundancy Analysis was intercepted because not all of the indicators are of the same scale. The Redundancy Analysis also shows that only 6.7% of the variance of service variables was explained by the Loyalty variable, while 27% of the variance of Loyalty variables was explained by the service variables.

Discriminant Analysis

This section uses all the variables from customers' attributes to which service they use to classify the customer churn or not. Firstly, I used the Stepwise Discriminant Analysis (SAS, The DISCRIM Procedure, n.d.) to remove the unimportant variables, that was PhoneService and OnlineBackup.

The variables selected in the above step were used for analysis. The test for homogeneity of covariance matrices was done. Because the Chi-Square value is smaller than 0.0001, we reject the hypothesis that the covariance matrices are homogeneous. Quadratic discriminant analysis is more appropriate than linear discriminant analysis.

Stepwise Selection Summary										
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	Contract_Month-to-month		0.1641	1382.34	<.0001	0.83589163	<.0001	0.16410837	<.0001
2	2	InternetService_Fiber optic		0.0556	414.68	<.0001	0.78939337	<.0001	0.21060663	<.0001
3	3	tenure		0.0390	285.86	<.0001	0.75858677	<.0001	0.24141323	<.0001
4	4	PaymentMethod_Electronic check		0.0164	117.07	<.0001	0.74617508	<.0001	0.25382492	<.0001
5	5	StreamingMovies		0.0081	57.14	<.0001	0.74016504	<.0001	0.25983496	<.0001
6	6	PaperlessBilling		0.0043	30.53	<.0001	0.73696736	<.0001	0.26303264	<.0001
7	7	SeniorCitizen		0.0031	22.18	<.0001	0.73465095	<.0001	0.26534905	<.0001
8	8	OnlineSecurity		0.0027	18.70	<.0001	0.73270257	<.0001	0.26729743	<.0001
9	9	InternetService_No		0.0039	27.38	<.0001	0.72986145	<.0001	0.27013855	<.0001
10	10	TechSupport		0.0046	32.18	<.0001	0.72653669	<.0001	0.27346331	<.0001
11	11	OnlineBackup		0.0018	12.45	0.0004	0.72525283	<.0001	0.27474717	<.0001
12	12	Contract_One year		0.0013	9.34	0.0023	0.72429057	<.0001	0.27570943	<.0001
13	13	StreamingTV		0.0009	6.57	0.0104	0.72361406	<.0001	0.27638594	<.0001
14	14	PhoneService		0.0009	6.10	0.0136	0.72298705	<.0001	0.27701295	<.0001
15	15	MultipleLines		0.0020	13.97	0.0002	0.72155232	<.0001	0.27844768	<.0001
16	16	Dependents		0.0005	3.81	0.0509	0.72116093	<.0001	0.27883907	<.0001
17	17	MonthlyCharges		0.0005	3.24	0.0721	0.72082891	<.0001	0.27917109	<.0001
18	16		PhoneService	0.0002	1.38	0.2399	0.72097067	<.0001	0.27902933	<.0001
19	15		OnlineBackup	0.0001	0.69	0.4046	0.72104195	<.0001	0.27895805	<.0001

Figure 9: Stepwise Discriminant Analysis

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
4452.540248	120	<.0001

Figure 10: Test of Homogeneity

Figure 11 shows that the quadratic discriminant analysis had a misclassification rate of 24.22%.

Number of Observations and Percent Classified into Churn			
From Churn	No	Yes	Total
No	3903 75.43	1271 24.57	5174 100.00
Yes	435 23.27	1434 76.73	1869 100.00
Total	4338 61.59	2705 38.41	7043 100.00
Priors	0.73463	0.26537	

Error Count Estimates for Churn			
	No	Yes	Total
Rate	0.2457	0.2327	0.2422
Priors	0.7346	0.2654	

Figure 11: Misclassification summary

Partial Least Squares Analysis

Partial least squares analysis was conducted to see how the variables are correlated with the Churn. PLS procedure (SAS, The PLS Procedure, n.d.) using leave one out cross-validation and van der Voet's T-square test for the number of factors.

Cross Validation for the Number of Extracted Factors			
Number of Extracted Factors	Root Mean PRESS	T**2	Prob > T**2
0	1.000142	601.5331	<.0001
1	0.878996	118.0803	<.0001
2	0.865048	49.46976	<.0001
3	0.858534	25.98504	<.0001
4	0.856449	17.71468	<.0001
5	0.85494	12.17006	0.0010
6	0.854241	9.93134	0.0020
7	0.853593	8.97534	0.0020
8	0.852618	4.874591	0.0290
9	0.852075	2.540388	0.1060
10	0.85164	0.507858	0.4710
11	0.851536	0.004305	0.9480
12	0.851531	0	1.0000
13	0.851542	0.176654	0.6600
14	0.851543	0.170015	0.6920
15	0.851545	0.214693	0.6450

Minimum root mean PRESS	0.8515
Minimizing number of factors	12
Smallest number of factors with p > 0.1	9

Figure 12: Cross Validation of Extracted Factors - Churn

The minimum PRESS occurs at 12 factors. However, the CVTEST option returns a nonsignificant T2 (where alpha=0.1) at 9 factors. Figure 13 shows that the variables that are important to classifying churn are having partners and dependents or not, tenure, streaming TV and movies, paperless billing, monthly charges, using Fiber Optic service, contract month-to-month and one year, and payment method using an electronic check.

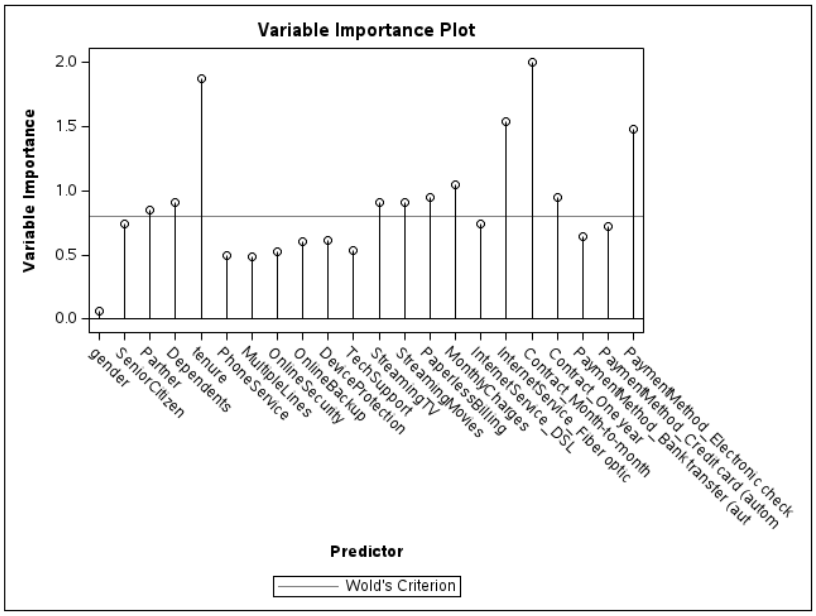

Figure 13: PLS Important Variables

The other analysis is between all variables and the tenure, which is the number of months people stay with the company. Figure 14 shows that the model required a minimum of ten extracted factors.

Cross Validation for the Number of Extracted Factors			
Number of Extracted Factors	Root Mean PRESS	T**2	Prob > T**2
0	1.000142	2209.369	<.0001
1	0.780207	708.5099	<.0001
2	0.70029	508.1149	<.0001
3	0.63731	187.4507	<.0001
4	0.615834	90.73664	<.0001
5	0.610444	60.74216	<.0001
6	0.608485	48.7856	<.0001
7	0.606564	35.53142	<.0001
8	0.602653	11.29673	<.0001
9	0.601556	3.423394	0.0600
10	0.601157	1.153043	0.2570
11	0.601061	0.060293	0.8060
12	0.601073	3.076584	0.0860
13	0.601057	0.94715	0.3200
14	0.601052	0.023683	0.8760
15	0.601052	0	1.0000

Minimum root mean PRESS	0.6011
Minimizing number of factors	15
Smallest number of factors with p > 0.1	10

Figure 14: Cross Validation of Extracted Factors – Tenure

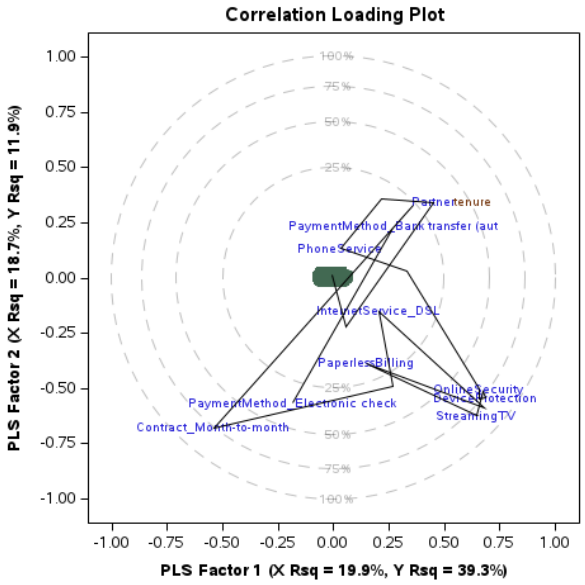

Figure 15: Correlation Loading Plot

The correlation loading plot in Figure 15 shows that tenure is positively correlated with Factor 1 and Factor 2. OnlineSecurity, DeviceProtection, and StreamingTV are most strongly correlated with Factor 1 and are well predicted by the factors.

4. Discussion

The Cluster Analysis shows the main group of customers is the group of young people, who choose Fiber Optic Internet service with month-to-month contracts and they do not often need Tech Support. The Correspondence Analysis explores the group of customer attributes to the services that telecom company provides. I found that the younger, having a partner but not having dependents would use streaming services, and being younger and having dependents would not use internet service. The Canonical Correlation Analysis was used to find the relationship between the group of services variable and the group of loyalty such as the churn and tenure. The discriminant analysis indicated that using all variables while removing the two unimportant variables, would be misclassified 24.22% of the total population. The partial least squares highlight the list of the important variables are having partners and dependents or not, tenure, streaming TV and movies, paperless billing, monthly charges, using Fiber Optic service, contract month-to-month and one-year, and payment method using electronic checks. The PLS model between all variables and tenure shows that OnlineSecurity, DeviceProtection, and StreamingTV are highly correlated with the tenure variable.

5. Conclusion

The main objective of this report was to explore the customer data of the telecom company from the customer attributes to the services they used. Several methods were used to explore the multivariate dataset, including Cluster Analysis, Correspondence Analysis, Canonical Correlation Analysis, Discriminant Analysis, and Partial least squares analysis.

6. Appendix

```
proc print DATA=data.churn(obs=10);
run;

data data.churn3;
set data.churn2;
        if Churn='No' then Churn_binary=0;
        else Churn_binary=1;
run;

proc standard data=data.churn3 mean=0 std=1 out=churn4;
run;

*Correspondence Analysis;
data data.addons;
        set data.churn;
        group=SeniorCitizen || Partner || Dependent || Gender;
run;

proc freq data=data.addons;
        tables group*StreamingMovies/ chisq;
run;

proc corresp data=data.Churn cross=rows observed rp short;
        tables   SeniorCitizen Partner Dependents Gender, StreamingTV;
run;
```

```
proc corresp data=data.Churn cross=rows observed rp short;
        tables   SeniorCitizen Partner Dependents Gender, StreamingMovies;
run;

*Distance;
proc distance data=data.churn method=DGOWER absent=0 out=data.gower;
        VAR interval(tenure MonthlyCharges) Nominal(gender--Dependents PhoneService--
PaymentMethod);
        id CustomerID;
run;

ods graphics on;
proc cluster data=data.gower method=CEN ccc pseudo print=15 outtree=tree;
id CustomerID;
run;

proc tree data=tree noprint n=7 out=out;
id CustomerID;
run;

proc sort data=data.churn;
by CustomerID;
run;

proc sort data=out;
by CustomerID;
run;
```

```
data clus;
   merge data.churn out;
   by CustomerID;
run;

*Plot;

proc sgplot data=clus;
vbar SeniorCitizen / group = Cluster groupdisplay = Cluster;
run;

proc sgplot data=clus;
vbar InternetService / group = Cluster groupdisplay = Cluster;
run;

proc sgplot data=clus;
vbar Contract / group = Cluster groupdisplay = Cluster;
run;

proc sgplot data=clus;
vbar TechSupport / group = Cluster groupdisplay = Cluster;
run;

proc sgplot data=clus;
vbar PhoneService / group = Cluster groupdisplay = Cluster;
run;
```

```
* Canonical Correlation Analysis;
proc cancorr data=data.churn3
    vprefix=S vname=Service
    wprefix=L wname=Loyalty;
        var PhoneService OnlineSecurity--StreamingMovies PhoneService InternetService_DSL-
-"InternetService_Fiber Optic"n;
  with tenure Churn_binary;
run;

proc cancorr data=data.churn3 out=churn_cancorr red
    vprefix=S vname=Service
        wprefix=L wname=Loyalty
        ncan=2;
        var PhoneService OnlineSecurity--StreamingMovies PhoneService InternetService_DSL-
-"InternetService_Fiber Optic"n;
        with tenure Churn_binary;
run;

* Canonical Discriminant Analysis;

proc stepdisc data=data.churn2 method=stepwise;
        class Churn;
        var gender--monthlycharges InternetService_DSL "InternetService_Fiber Optic"n
"Contract_Month-to-month"n
        "Contract_One year"n "PaymentMethod_Bank transfer (aut"n--
"PaymentMethod_Electronic check"n;
run;
```

```
proc discrim data=data.churn2 pool=test slpool=.05;
        class Churn;
        priors prop;
        var &_STDVAR;
run;


*partial;
proc pls data = data.churn3 method = pls(algorithm=nipals)
cv=one cvtest(seed=608789001)
plot=(vip xyscores xscores parmprofiles dmod);
model tenure = gender--Dependents PhoneService--MonthlyCharges InternetService_DSL
"InternetService_Fiber Optic"n "Contract_Month-to-month"n
        "Contract_One year"n "PaymentMethod_Bank transfer (aut"n--
"PaymentMethod_Electronic check"n;
run;


proc pls data = data.churn3 method = pls(algorithm=nipals)
cv=one cvtest(seed=608789001)
plot=(vip xyscores xscores parmprofiles dmod);
model Churn_binary = gender--MonthlyCharges InternetService_DSL "InternetService_Fiber
Optic"n "Contract_Month-to-month"n
        "Contract_One year"n "PaymentMethod_Bank transfer (aut"n--
"PaymentMethod_Electronic check"n;;
run;
```

Bibliography

SAS. (n.d.). *The CLUSTER Procedure*. From https://documentation.sas.com/doc/en/statug/15.2/statug_cluster_details01.htm

SAS. (n.d.). *The DISTANCE Procedure*. From https://documentation.sas.com/doc/en/statug/15.2/statug_distance_syntax01.htm#statug.distance.methodopt

Caliński, T. (1974). A Dendrite Method for Cluster Analysis.

Duda, R. (1974). Pattern classification and scene analysis.

IBM. (2024). *Cramér's V*. From https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=terms-cramrs-v

Kaggle. (n.d.). *Telco Customer Churn*. From https://www.kaggle.com/datasets/blastchar/telco-customer-churn/data

SAS. (n.d.). *The CORRESP Procedure*. From https://documentation.sas.com/doc/ko/pgmsascdc/9.4_3.4/statug/statug_corresp_toc.htm

SAS. (n.d.). *The CANCORR Procedure*. From https://documentation.sas.com/doc/ko/pgmsascdc/9.4_3.4/statug/statug_cancorr_toc.htm

SAS. (n.d.). *The DISCRIM Procedure*. From https://documentation.sas.com/doc/ko/pgmsascdc/9.4_3.4/statug/statug_discrim_toc.htm

SAS. (n.d.). *The PLS Procedure*. From https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_pls_syntax01.htm#statug.pls.pls_prc_methodopt