Canonical Correlation Analysis for Credit Card Payment Method Prediction Data Set

ST405 Mini Project 02

- Agamini Maheesha Weerakkody -

S/18/834

1 Introduction

Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a statistical method used to explore the relationships between two sets of variables. It identifies and measures the associations between these sets by finding pairs of linear combinations that have the highest possible correlation with each other. This technique is particularly useful in fields like psychology, finance, and bioinformatics, where understanding the interplay between multiple variables is crucial for drawing meaningful conclusions from complex data sets.

Credit Card Payment Method Prediction Data Set

The "Credit Card Payment Method Prediction Data Set" is an advanced-level dataset comprising 30,000 rows and 24 columns, designed for estimating the probability of default payment by credit card clients. Sourced from the Machine Learning Repository of the University of California, Irvine, this dataset is ideal for practicing exploratory data analysis, data visualization, and classification modeling techniques. It includes detailed information on various attributes such as credit amount, demographics, payment history, bill statements, and previous payments, allowing for comprehensive analysis using both supervised and unsupervised learning methods.

1. **LIMIT_BAL:** Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
2. **SEX:** Gender (1 = male; 2 = female).
3. **EDUCATION:** Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
4. **MARRIAGE:** Marital status (1 = married; 2 = single; 3 = others).
5. **AGE:** Age (year).
6. **PAY_0:** History of past payment. The repayment status in September, 2005.
7. **PAY_2:** History of past payment. The repayment status in August, 2005.
8. **PAY_3:** History of past payment. The repayment status in July, 2005.
9. **PAY_4:** History of past payment. The repayment status in June, 2005.
10. **PAY_5:** History of past payment. The repayment status in May, 2005.
11. **PAY_6:** History of past payment. The repayment status in April, 2005.
12. **BILL_AMT1:** Amount of bill statement in September, 2005 (NT dollar).
13. **BILL_AMT2:** Amount of bill statement in August, 2005 (NT dollar).
14. **BILL_AMT3:** Amount of bill statement in July, 2005 (NT dollar).
15. **BILL_AMT4:** Amount of bill statement in June, 2005 (NT dollar).
16. **BILL_AMT5:** Amount of bill statement in May, 2005 (NT dollar).
17. **BILL_AMT6:** Amount of bill statement in April, 2005 (NT dollar).

18. **PAY_AMT1:** Amount of previous payment. Paid in September, 2005 (NT dollar).
19. **PAY_AMT2:** Amount of previous payment. Paid in August, 2005 (NT dollar).
20. **PAY_AMT3:** Amount of previous payment. Paid in July, 2005 (NT dollar).
21. **PAY_AMT4:** Amount of previous payment. Paid in June, 2005 (NT dollar).
22. **PAY_AMT5:** Amount of previous payment. Paid in May, 2005 (NT dollar).
23. **PAY_AMT6:** Amount of previous payment. Paid in April, 2005 (NT dollar).
24. **Default Payment Next Month:** Probability of Default. (1: Yes, 0: No).

2 Methodology

For this study, I will apply canonical correlation analysis (CCA) to explore the relationships between two subsets of the data: bill data and payment data. The bill data subset includes the variables BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, and BILL_AMT6, representing the amounts of bill statements over six months. The payment data subset includes the variables PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, and PAY_AMT6, representing the amounts of previous payments over the same period. CCA will help identify how these two sets of variables are related by finding the pairs of linear combinations that have the highest correlations, providing insights into the factors that influence the likelihood of credit card default.

3 Results and Discussion

Split the dataset into two sets. Set 1 is “bill_data” and Set 2 is “payment_data”. There are 6 variables in each set. We can conclude that there are 6 canonical covariate pairs.

- Fit the canonical correlation model and get canonical correlations. There are 6 (equal to number of variables in Set 1(bill_data); small set) canonical correlations in this model

```
0.7169289 0.6136829 0.5528704 0.5337309 0.4043431 0.1621175
```

We can see the first two canonical correlations have high correlation therefore first and second cano nical covariate pairs are highly correlated. Second two canonical covariate pairs are moderately corr elated but fourth and fifth canonical covariate pairs are poorly correlated.

- Test for independence between canonical covariate pairs

```
Test of H0: The canonical correlations in the
current row and all that follow are zero

     CanR LR test stat approx F numDF  denDF   Pr(> F)
1 0.71693      0.12254  2242.21    36 131689 < 2.2e-16 ***
2 0.61368      0.25213  2001.00    25 111406 < 2.2e-16 ***
3 0.55287      0.40444  1974.95    16  91622 < 2.2e-16 ***
4 0.53373      0.58249  2016.56     9  72990 < 2.2e-16 ***
5 0.40434      0.81452  1619.91     4  59984 < 2.2e-16 ***
6 0.16212      0.97372   809.56     1  29993 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

All canonical covariate pairs are significantly correlated. Therefore, we can conclude that two sets of variables are dependent on one another at 1% significance level.

- Significant canonical correlations and squared canonical correlations.

  All 6 canonical covariate pairs are significant

```
0.51398704 0.37660674 0.30566563 0.28486867 0.16349338 0.02628209
```

  From squared canonical correlation we can conclude that, 51.39% of the variation in first canonical variable of “bill_data” set is explained by the variation in first canonical variable of “payment_data” set. 37.66% of the variation in the second canonical variable of “bill_data” set is explained by the variation in second canonical variable of “payment_data” set. 30.56% of the variation in the third canonical variable of “bill_data” set is explained by the variation in third canonical variable of “payment_data” set. Others have low values. Therefore, only the first three canonical correlations are important.

- Estimated canonical coefficients for the “bill_data” set

	1	2	3	4	5	6
BILL_AMT1	-0.363650	-0.130814	-1.639133	2.0398007	0.738498	1.76350396
BILL_AMT2	2.2296050	0.0945234	1.321579	-3.0458209	0.018576	-0.1083793
BILL_AMT3	-3.168819	0.5370346	0.454330	-0.4539763	0.004448	0.09385325
BILL_AMT4	1.1379350	-1.877875	2.009942	2.0027430	-0.60671	-0.2189401
BILL_AMT5	0.5303220	3.4068343	-0.764868	0.2718003	-0.78813	-0.3857955
BILL_AMT6	-0.421722	-2.467751	-1.496706	-0.9308426	-0.20844	-0.5379414

  The magnitudes of the coefficients give the contribution of the individual variables to the corresponding canonical variable. ‘BILL_AMT2’ gives the best contribution to the first canonical variable of “payment_data” set. ‘BILL_AMT5’ gives the highest contribution for the second canonical variable. ‘BILL_AMT4’ gives the highest contribution for the second canonical variable.

- Estimated canonical coefficients for the “payment_data” set

	1	2	3	4	5	6
PAY_AMT1	0.292129	0.034886	0.5457156	-08270403	-0.300152	-0.1064727
PAY_AMT2	-0.96562	0.113727	0.2509522	0.2321429	-0.255802	-0.1821586
PAY_AMT3	0.517025	-0.288871	0.1204097	0.6843822	-0.524819	-0.200867
PAY_AMT4	0.048011	0.6363077	-0.7438518	-0.160642	-0.351471	-0.0837535
PAY_AMT5	-0.169771	-0.820648	-0.4747600	-0.307603	-0.125186	0.1803702
PAY_AMT6	0.013077	0.1923429	0.1710883	0.1632821	-0.052198	0.9925306

‘PAY_AMT2’ give the best contribution to the first canonical variable of “payment_data” set. ‘PAY_AMT5’ give the best contribution to the second canonical variable of “payment_data” set. ‘PAY_AMT4’ give the best contribution to the third canonical variable of “payment_data” set. However, all of them are negative contributions.

- The correlation between the “bill_data” variables and the canonical variables for “bill_data” set

	1	2	3	4	5	6
BILL_AMT1	0.01080102	-0.0539787	-0.0831876	-0.6202416	-0.5830855	0.8038742
BILL_AMT2	0.06274956	-0.0433795	007530774	-0.2794038	-0.6670724	0.6822448
BILL_AMT3	-0.2631861	-0.0263949	0.12250568	-0.1650077	-0.7544592	0.5644399
BILL_AMT4	0.00571437	-0.1164007	0.13159272	0.03662009	-0.8794679	0.4407910
BILL_AMT5	0.01536802	-0.0876616	-0.1136960	-0.0535341	-0.9230678	0.3524663
BILL_AMT6	-0.0364972	-0.1883145	-0.2385269	-0.1523384	-0.8887936	0.3052040

Only the ‘BILL_AMT3’ variable has a significant correlation with first canonical covariate of Set1. Therefore, the first canonical covariate of Set1 is the measure of ‘BILL_AMT3’ variable. Rest of the correlations between first canonical covariate and variables of Set1 is very low. ‘BILL_AMT4’ and ‘BILL_AMT6’ have weak negative correlation with the second canonical covariate of Set1. ‘BILL_AMT3’, ‘BILL_AMT4’, ‘BILL_AMT5’ and ‘BILL_AMT6’ have weak correlations with the third canonical covariate.

- The correlation between the “payment_data” variables and the canonical variables for “payment_data” set

	1	2	3	4	5	6
PAY_AMT1	0.13356519	0.03288293	0.4606007	-0.6355476	-0.6039769	-0.0147378
PAY_AMT2	-0.7756536	0.04702587	0.2433767	0.14046351	-0.5641569	-0.0877286
PAY_AMT3	0.33982591	-0.2166498	0.1108009	0.47547854	-0.7675859	-0.1001816
PAY_AMT4	0.02052580	0.49650287	-0.6087836	-0.1567575	-0.5982186	0.00277912
PAY_AMT5	-0.2094575	-0.7194179	-0.4156119	-0.2785243	-0.3610311	0.24065092
PAY_AMT6	-0.0194574	001409477	0.1406577	0.08464035	-0.3085441	0.92607078

‘PAY_AMT2’ variable has a high and negative correlation with first canonical covariate of Set2. All the other correlations between variables and first canonical variable have weak correlation. Therefore, first canonical covariate of Set2 is a measure of ‘PAY_AMT2’. ‘PAY_AMT5’ variable is highly negatively correlated with the second canonical covariate of Set2 and ‘PAY_AMT4’ is moderately correlated with third canonical covariate of set2.

- The correlation between the “bill_data” variables and the canonical variables for “payment_data” set

	1	2	3	4	5	6
BILL_AMT1	0.00774356	-0.3312582	-0.0459919	-0.0331042	-0.2365766	0.13032210
BILL_AMT2	0.04498295	-0.0266212	0.04163542	-0.1491264	-0.2697262	0.11060384
BILL_AMT3	-0.1886857	-0.0161981	0.06772976	-0.0880697	-0.3050601	0.09150560
BILL_AMT4	0.00409679	-0.0714331	0.07275371	0.01954527	-0.3556068	0.07145995
BILL_AMT5	0.01101778	0.05379648	-0.0628591	-0.0285727	-0.3732361	0.05714096
BILL_AMT6	-0.0261659	-0.1155654	-0.1318744	-0.0813077	-0.3593776	0.0494782

Only Variable ‘BILL_AMT3’ has a significant correlation with first canonical covariate of Set2 and it is negative weak correlation of -0.1886857. All the other variables do not have significant correlation with first canonical covariate of Set2. There is a weak negative correlation between variable ‘BILL_AMT6’ and second canonical covariate and the third variable canonical covariate of Set2.

- The correlation between the “payment_data” variables and the canonical variables of “bill_data” set

	1	2	3	4	5	6
PAY_AMT1	0.09575675	0.02017969	0.25465249	-0.3392113	-0.2442139	-0.0023892
PAY_AMT2	-0.5560884	0.02885897	0.13455579	0.05584701	-0.2281130	-0.0142223
PAY_AMT3	0.23436101	-0.1329543	0.06125851	0.25377758	-0.3103681	-0.0162411
PAY_AMT4	0.01471554	0.30469534	-0.3365784	-0.0836663	-0.2418856	0.00045054
PAY_AMT5	-0.1501662	-0.4414954	-0.2297794	-0.1486571	-0.1459804	0.03901373
PAY_AMT6	-0.0139495	0.08649723	0.07776550	0.04517517	-0.1247577	0.15013229

There is a negative moderate correlation between variable ‘PAY_AMT2’ and first canonical covariate of Set1.’PAY_AMY3’ is weakly positively correlated with first canonical covariate of Set1. There is a moderate negative correlation between variable ‘PAY_AMT5’ and second canonical covariate of Set1. There is a positive moderate correlation with ‘PAY_AMT4’ and second canonical covariate of set1. ‘PAY_AMT4’ is weakly negatively correlated with third canonical covariate of Set1. ‘PAY_AMT1’, ‘PAY_AMT2’ and ‘PAY_AMT5’ are weakly correlated with third canonical covariate of Set1.

4 Conclusion and Recommendation

- The canonical correlation analysis reveals significant correlations between the "bill_data" and "payment_data" sets.

- The first three canonical correlations are highly significant, explaining 51.39%, 37.66%, and 30.56% of the variation in the "bill_data" and "payment_data" sets, respectively.

- The canonical coefficients show that 'BILL_AMT2' and 'PAY_AMT2' are the most contributing variables to the first canonical variables of their respective sets.

- The correlations between the variables and canonical variables reveal that only a few variables have significant correlations with the canonical variables, indicating that the canonical variables are not strongly related to all individual variables.

- The results provide insights into the relationships between the "bill_data" and "payment_data" sets, highlighting the importance of 'BILL_AMT2', 'PAY_AMT2', and other variables in explaining the correlations between the two sets.

5 References

Canonical Correlation Analysis
https://www.wallstreetmojo.com/canonical-correlation-analysis/

Canonical Correlation in R using mtcars
https://medium.com/@josef.waples/canonical-correlation-analysis-cca-in-r-rstudio-using-mtcars-1669b1c56731

Canonical Correlation Analysis STATA Data Analysis Example
https://stats.oarc.ucla.edu/stata/dae/canonical-correlation-analysis/

6 Appendix

Data Set

ID <dbl>	LIMIT_BAL <dbl>	SEX <dbl>	EDUCATION <dbl>	MARRIAGE <dbl>	AGE <dbl>	PAY_0 <dbl>	PAY_2 <dbl>	PAY_3 <dbl>	PAY_4 <dbl>	PAY_5 <dbl>	PAY_6 <dbl>	BILL_AMT1 <dbl>	BILL_AMT2 <dbl>	BILL_AMT3 <dbl>	BILL_AMT4 <dbl>	BILL_AMT5 <dbl>	BILL_AMT6 <dbl>	PAY_AMT1 <dbl>
1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0
2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0
3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518
4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000
5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000
6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500

6 rows | 1-19 of 25 columns

```
library(tidyverse)
library(CCA)
library(CCP)
library(candisc)
library(skimr)
credit_card_clients<-read_csv("../Data/credit_card_clients.csv")
skim(credit_card_clients)
head(credit_card_clients)
credit_card_clients<-apply(credit_card_clients,2,scale)
# Subset the data into two sets
bill_data <- credit_card_clients[, c("BILL_AMT1", "BILL_AMT2", "BILL_AMT3", "BILL_AMT4", "BILL_AMT5", "BILL_AMT6")]
payment_data <- credit_card_clients[, c("PAY_AMT1", "PAY_AMT2", "PAY_AMT3", "PAY_AMT4", "PAY_AMT5", "PAY_AMT6")]
matcor(bill_data,payment_data)
```

```
# Perform canonical correlation analysis
cc_model <- cc(bill_data, payment_data)
cc_model$cor
Wilks(cancor(bill_data,payment_data))
cc_model$cor^2
cc_model$xcoef
cc_model$ycoef
loadings<-comput(bill_data,payment_data,cc_model)
loadings$corr.X.xscores
loadings$corr.Y.yscores
loadings$corr.X.yscores
loadings$corr.Y.xscores
```