Maestría en Inteligencia Artificial Aplicada

Inteligencia artificial y aprendizaje automático

Dalina Aidee Villa Ocelotl (A01793258)
Miguel Guillermo Galindo Orozco (A01793695)

Actividad semanal 4

Análisis de Componentes Principales (PCA)

Profesor:Jobish Vallikavungal

10 de octubre de 2022

PCA - Ejemplo de aplicación

Introducción

El análisis de componentes principales se identifica como una técnica de reducción de dimensión que permite pasar de una gran cantidad de variables interrelacionadas a unas pocas con componentes principales que recopilan la información de las variables explícitas.

Las componentes principales van de acuerdo a la siguiente definición donde las variables aleatorias son las variables de nuestra base de datos de tarjetas de crédito.

Definición *Se definen las p componentes principales de X como las variables aleatorias $(Z_1, \ldots, Z_p)$ tales que*

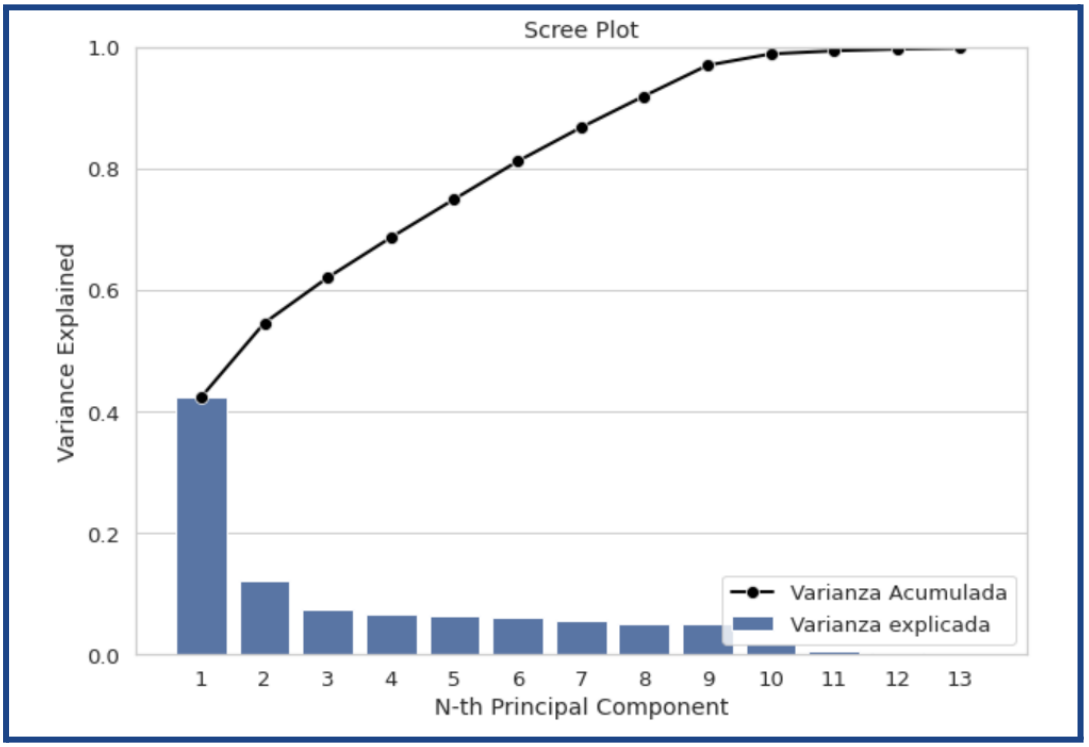
$$Z_1 = v_1'X, \ldots, Z_p = v_p'X \qquad v_1, \ldots, v_p \in \mathbb{R}^p$$

$$\begin{aligned}
Var(Z_1) &= \max\{Var(v'X) : v \in \mathbb{R}^p,\ v'v = 1\} \\
Var(Z_2) &= \max\{Var(v'X) : v \in \mathbb{R}^p,\ v'v = 1,\ v_1'v = 0\} \\
&\vdots \quad\quad \vdots \\
Var(Z_j) &= \max\{Var(v'X) : v \in \mathbb{R}^p,\ v'v = 1,\ v_1'v = 0, \ldots,\ v_{j-1}'v = 0\}\} \\
&\vdots \quad\quad \vdots \\
Var(Z_p) &= \max\{Var(v'X) : v \in \mathbb{R}^p,\ v'v = 1,\ v_1'v = 0, \ldots,\ v_{p-1}'v = 0\}\}
\end{aligned}$$

Para el presente ejercicio se utilizó la base sobre personas con crédito, dicha base se consume después de la limpieza realizada en ejercicios anteriores (Semana 3 Actividad 1).

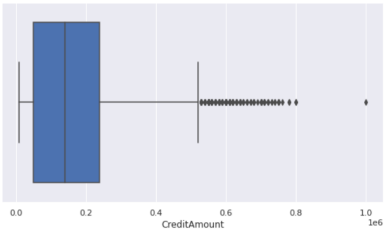
Primeramente se obtuvo la base de datos limpia y con datos normalizados lo cuales nos proporciona consistencia en el análisis, después se aplicó el análisis de PCA a la base estandarizada primeramente a todas las variables para después ver hasta qué número de variable se tomaron los componentes, después se validó el comportamiento de cada variable y su relevancia dentro de las características, y finalmente se analizó si esa importancia es asociada a los valores atípicos de cada variable.

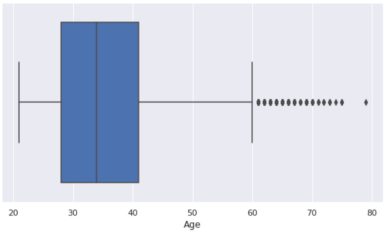
Captura de resultados.

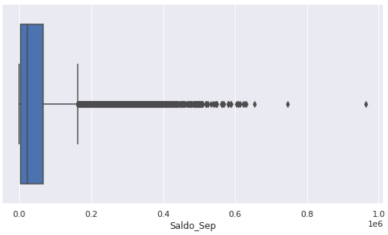


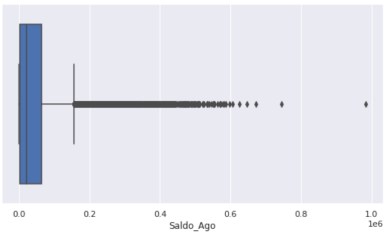
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Desviación_Estándar	2.43	1.31	1.02	0.96	0.94	0.93	0.88	0.85	0.84	0.51	0.27	0.2	0.16
Prop_Varianza_explicada	0.42	0.12	0.07	0.07	0.06	0.06	0.06	0.05	0.05	0.02	0.01	0.0	0.00
Acum_Prop_Varianza_explicada	0.42	0.55	0.62	0.69	0.75	0.81	0.87	0.92	0.97	0.99	0.99	1.0	1.00

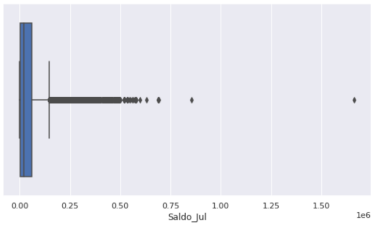
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
CreditAmount	0.1655	0.3008	-0.3786	-0.1998	-0.0303	-0.0810	0.1111	-0.0547	-0.8216	-0.0295	-0.0063	0.0156	-0.0007
Age	0.0327	0.0720	-0.8700	0.3370	0.0340	0.0731	-0.0784	0.0311	0.3311	-0.0090	0.0003	-0.0013	0.0002
Saldo_Sep	0.3724	-0.1910	-0.0334	-0.0613	-0.0386	-0.0448	0.0073	0.0060	0.0089	0.5667	0.4171	-0.4331	-0.1878
Saldo_Ago	0.3831	-0.1752	-0.0016	0.0089	-0.0807	-0.0323	-0.0322	-0.1343	0.0170	0.3869	0.0383	0.3475	0.3374
Saldo_Jul	0.3881	-0.1274	0.0347	0.0618	-0.1190	0.0920	-0.1184	0.0930	-0.0198	0.1228	-0.4858	0.4942	-0.0956
Saldo_Jun	0.3915	-0.1204	0.0339	0.0744	-0.0254	0.0109	0.1244	0.0398	0.0188	-0.2041	-0.5252	-0.4942	-0.3501
Saldo_May	0.3883	-0.1065	0.0334	0.0383	0.1139	-0.0904	-0.0075	0.0497	0.0232	-0.4203	0.0736	-0.2441	0.7161
Saldo_Abr	0.3809	-0.0936	0.0177	-0.0738	0.1545	0.0766	0.0083	0.0023	0.0607	-0.4897	0.5100	0.3370	-0.4324
Pagos_Sep	0.1353	0.3830	0.1730	0.3619	-0.2300	0.0253	-0.2018	-0.7483	0.0285	-0.0601	0.0461	-0.0676	-0.0455
Pagos_Ago	0.1169	0.4084	0.2003	0.3468	-0.1759	0.3975	-0.2781	0.5777	-0.1161	0.0494	0.1473	-0.0686	0.0398
Pagos_Jul	0.1282	0.3924	0.1219	0.2448	0.2475	-0.0918	0.7854	0.0682	0.1522	0.1441	0.0017	0.1232	0.0233
Pagos_Jun	0.1172	0.3493	0.0625	-0.0934	0.6107	-0.4601	-0.4637	0.0775	0.0970	0.1257	-0.1132	0.0028	-0.0780
Pagos_May	0.1139	0.3042	-0.0602	-0.6100	0.1526	0.6143	0.0137	-0.1622	0.2540	0.0607	-0.0986	-0.0684	0.0945
Pagos_Abr	0.1060	0.3230	-0.0493	-0.3667	-0.6299	-0.4542	0.0259	0.1861	0.3142	-0.0930	0.0302	0.0246	-0.0142

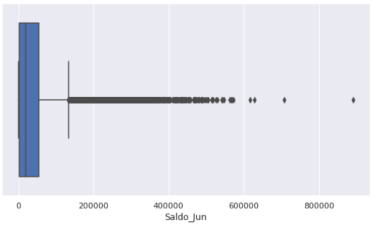


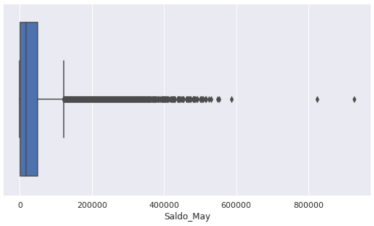


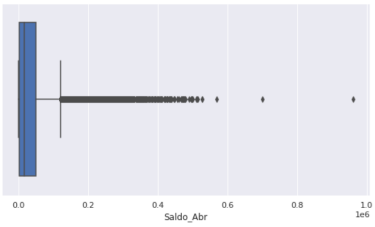


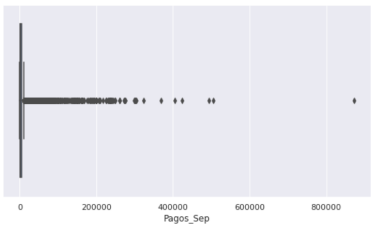


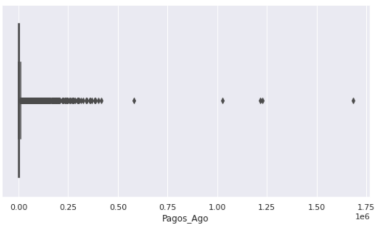


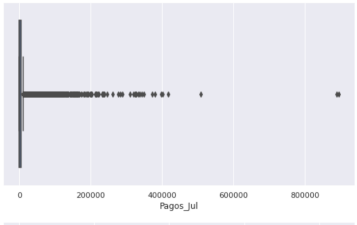


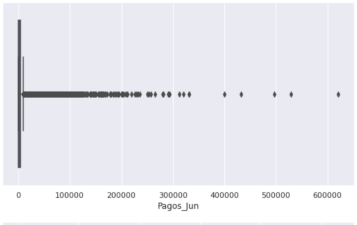


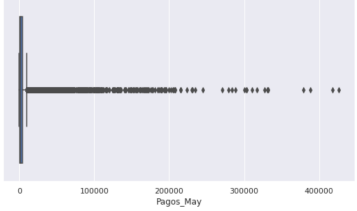


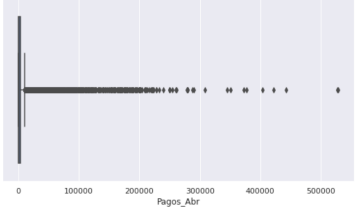












A continuación respondemos las preguntas de la parte 2, considerando que los comentarios por cada resultado se encuentren dentro del código mismo.

¿Cuál es el número de componentes mínimo y por qué?

Con los resultados anteriores, de la tabla (PCA_resumen) y la gráfica (Scree Plot), concluimos que es pertinente utilizar los primeros 7 Componentes Principales, ya que con esto capturamos el 87% de la varianza del total de variables numéricas.

Lo anterior lo definimos como mínimo, ya que por la definición de la metodología de PCA, el objetivo es capturar la mayor cantidad de varianza para garantizar la explicabilidad de la información con la menor cantidad de variables posible. Como ejemplo, si sólo utilizamos el primer componente nos quedamos con el 42% de la varianza, la cual no sería suficiente para capturar la mayor cantidad de información de todas las variables numéricas.

¿Cuál es la variación de los datos que representan esos componentes?

Para la aplicación en el ejercicio se establece que hasta hasta el séptimo componente con lo que se recoge el 87 % de la varianza.

En otras palabras, transformando las variables a estos componentes (PC1, PC2, PC3, PC4, PC5, PC6, PC7) y sólo utilizando esta información, reducimos de 14 a 7 variables (50% de las numéricas), capturando el 87% de la información.

No utilizamos más componentes porque cada componente extra aporta a más 5% extra de información, aumentando la dimensionalidad por lo que no es un lift significativo en la información a utilizar.

¿Cuál es la pérdida de información después de realizar PCA?

La pérdida de información al utilizar el análisis de componentes principales es del 13% considerando que sólo conservamos 7 componentes donde se captura el 87% de la varianza.

Por otro lado, observamos que los componentes principales no capturan información sobre la edad, por el tipo de variable se esperaba que los componentes seleccionados describiera mejor dicha información. Se decide que se evaluará después si se aplicará directamente. Esto también aplicaría para el monto del crédito.

De las variables originales, ¿Cuál tiene mayor y cuál tiene menor importancia en los componentes principales?

En los primeros componentes se observa que tienen mayor importancia las variables de saldos dependiendo de la temporalidad. Por otro lado, los siguientes componentes capturan información sobre pagos y comportamiento crediticio del cliente. Finalmente, los tributos específicos del cliente como edad, y el monto solicitado son considerados dentro de los componentes seleccionados.Sin embargo, no figuran como las variables de mayor peso dentro de los mismos. Se considera utilizarlas directamente, ya que bajo nuestra experiencia consideramos que dichas variables aportan mucha información al momento de decidir otorgar el préstamo, y que esto esté asociado a su forma de pago.

¿Cuándo se recomienda realizar un PCA y qué beneficios ofrece para Machine Learning?

Se recomienda utilizar PCA cuando se busca optimizar recursos computacionales y de análisis para grandes bases de datos. Trae beneficios tales como el mejor uso de recursos, capturar información, reducir el ruido de variables que no aportan información y garantizar que se utiliza información sobre la tendencia en los datos observados.

Además, se tiene mayor precisión en la predicción y se reduce el tiempo disminuyendo las características. En el presente caso vemos que se requiere de al menos 7 componentes para obtener la mayor varianza, lo que nos indica que igual pudiera solo utilizarse las variables Edad, Monto de Crédito y los primeros componentes.

Se recomienda utilizar PCA por estas dos ventajas muy importantes, como lo son eliminar dimensiones y extraer el menor número de características con la mayor cantidad de información posible.

Referencias

- (S/f). Usc.es. Recuperado el 10 de octubre de 2022, de http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP/MATERIALESMASTER/Mat_14_master0809multi-tema5.pdf
- (S/f-b). Uc3m.es. Recuperado el 10 de octubre de 2022, de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema3am.pdf