Linear vs PCR vs PLS models

Modelado Predictivo

ITESO, Universidad Jesuita de Guadalajara

Figure 1: nothing

Gregorio Alvarez

Introducción:

Este informe tiene como objetivo comparar y analizar las diferencias entre las predicciones realizadas por diferentes modelos en un estudio de regresión. En particular, se examinarán las predicciones generadas por un modelo lineal estándar, un modelo lineal después de la selección de variables, un modelo lineal después de aplicar PCA a las variables de entrada y un modelo PLS (Partial Least Squares). El objetivo es determinar cuál de estos enfoques ofrece las predicciones más precisas y confiables al ser utilizados en la base de datos abalone.

Descripción de la base de datos

La base de datos Abalone contiene informacion de los abalones, un tipo de molusco. En este caso se busca predecir la edad del abalon dado el resto de

variables. La base de datos contiene 4177 instancias y 9 atributos. Ya que en este estudio no se realizaran transformaciones de variables categoricas, estas no seran tomadas en cuenta en los modelos.

A continuación se muestra una descripción de cada columna por fila:

Variables independientes:

1. Length: Continuo, longitud más larga medida en mm.
2. Diam: Continuo, diámetro perpendicular a la longitud en mm.
3. Height: Continuo, altura con respecto al cono en mm.
4. Whole: Continuo, peso entero del abalón en gramos.
5. Shucked: Continuo, peso del abalón sin la concha en gramos.
6. Visceras: Continuo, peso de las vísceras en gramos.
7. Shell: Continuo, peso del caparazón en gramos.

Variable dependiente:

1. Rings: Entero, edad del abalón en años.

Metodología

Preprocesamiento:

Se realizó un análisis exploratorio de los datos para entender su estructura y características. Esto incluyó la identificación de variables y la comprobación de la existencia de valores atípicos.

	Length	Diam	Height	Whole	Shucked	Viscera	Shell	Rings
mean	0.523992	0.407881	0.139516	0.828742	0.359367	0.180594	0.238831	9.933684
std	0.120093	0.099240	0.041827	0.490389	0.221963	0.109614	0.139203	3.224169
min	0.075000	0.055000	0.000000	0.002000	0.001000	0.000500	0.001500	1.000000
max	0.815000	0.650000	1.130000	2.825500	1.488000	0.760000	1.005000	29.000000

Figure 2: “Tabla estadisticos”

Se creó una gráfica de pares para visualizar la relación entre la variable dependiente e independiente. Se encontraron valores atípicos en la variable de altura

Se decidió remover los valores atípicos en base a su z-score. Se eliminaron los datos con z-score mayor a 3 o menor a -3.

Regresión Lineal:

Se procedió a separar los datos en conjuntos de entrenamiento y prueba. Se utilizó una proporción de 70% para entrenamiento y 30% para prueba.

Utilizando los datos de entrenamiento, se creó un modelo lineal.

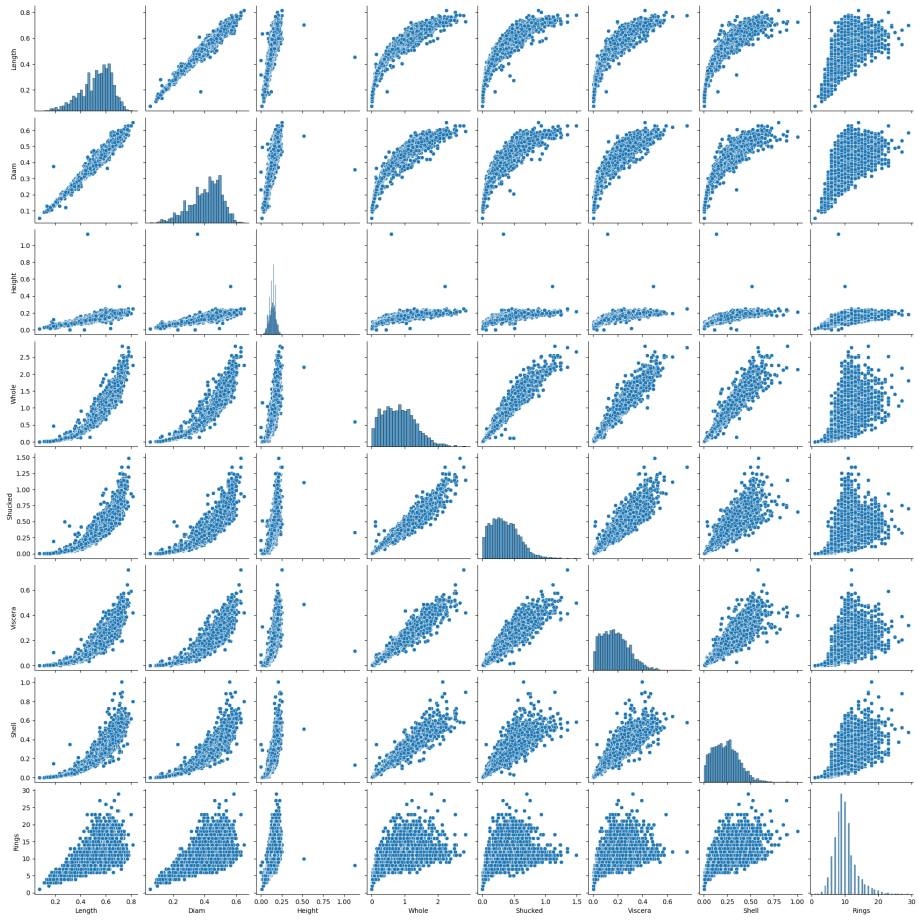

Figure 3: “pairplot”

Se evaluó el modelo utilizando los datos de prueba y entrenamiento. Como métricas se utilizó el error cuadrático medio (RMSE) y el coeficiente de determinación (R^2).

Se obtuvieron los siguientes resultados:

set	RMSE	R^2
Test	2.1167	0.5684
Train	2.2211	0.5244

R.L. con Selección de variables

se llevó a cabo una selección de variables basada en su correlación. Para ello, se realizó un análisis de correlación utilizando una tabla de correlación, donde se determinó que la correlación entre la variable dependiente y las variables independientes era baja.

	Length	Diam	Height	Whole	Shucked	Viscera	Shell	Rings
Length	1.000000	0.986756	0.900583	0.925684	0.898469	0.903367	0.898655	0.555530
Diam	0.986756	1.000000	0.906923	0.925781	0.893588	0.899971	0.906263	0.573586
Height	0.900583	0.906923	1.000000	0.889000	0.837582	0.866889	0.892005	0.609258
Whole	0.925684	0.925781	0.889000	1.000000	0.969350	0.966267	0.955924	0.540181
Shucked	0.898469	0.893588	0.837582	0.969350	1.000000	0.931788	0.883118	0.420601
Viscera	0.903367	0.899971	0.866889	0.966267	0.931788	1.000000	0.908072	0.503512
Shell	0.898655	0.906263	0.892005	0.955924	0.883118	0.908072	1.000000	0.627833
Rings	0.555530	0.573586	0.609258	0.540181	0.420601	0.503512	0.627833	1.000000

Figure 4: “abs_cor”

Sin embargo, para obtener una visión más completa de las relaciones entre las variables, se utilizó un enfoque gráfico basado en la técnica de clusterización. Mediante este análisis, se pudo observar que todas las variables presentaban una correlación significativa entre sí. Además, se identificaron tres grupos de variables que mostraban la mayor correlación dentro de cada grupo.

Finalmente, se seleccionaron las variables en base a un criterio de correlación global y correlación promedio entre pares, con lo que se obtuvieron las variables Length, Height, Shucked, Shell.

Se creó un modelo lineal y se evaluaron los resultado de las predicciones, los cuales se muestran a continuación.

set	RMSE	R^2
Test	2.1809	0.5418
Train	2.2638	0.5059

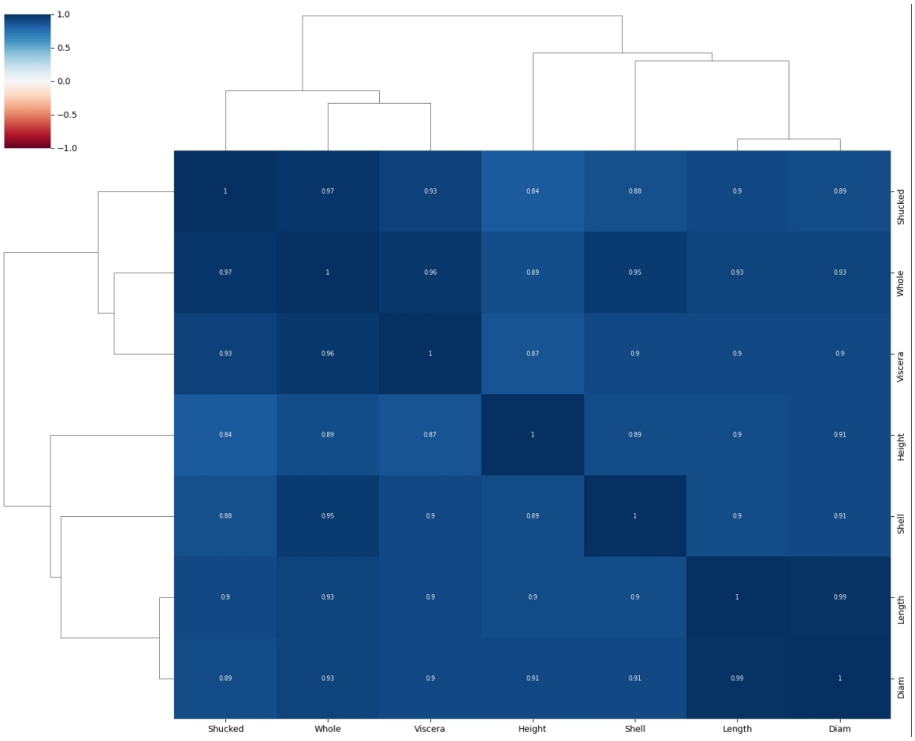

Figure 5: “clus_cor”

PCR

Se aplico PCA a las variables de entrada. Una vez aplicado el PCA, se obtuvo un gráfico de varianza acumulada. Basandose en el criterio del codo, se optó por seleccionar las primeras 3 componentes principales.

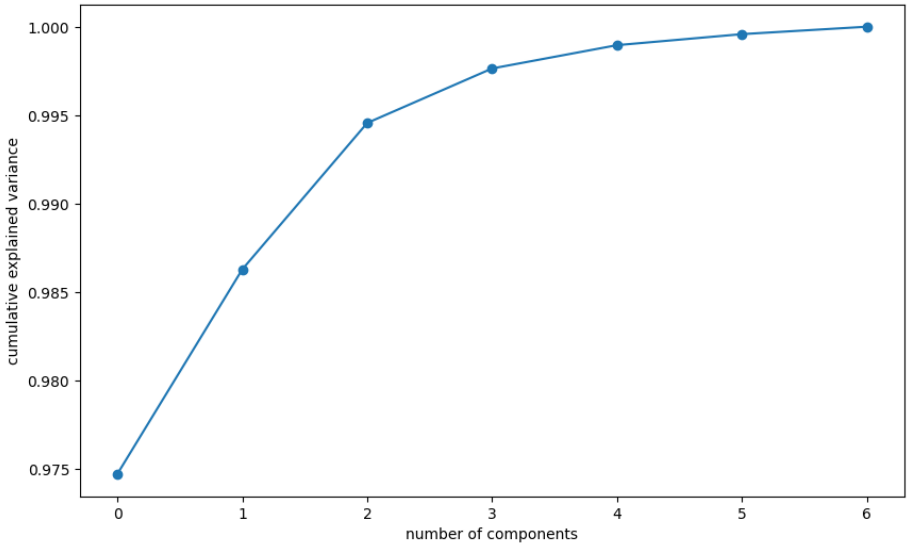

Figure 6: cumula_var

Se realizó un análisis gráfico para evaluar la relación entre las variables seleccionadas y la variable dependiente. Tras analizar el gráfico, se encontró que la relación lineal se mantenía solo para las primeras dos variables principales.

Con las variables seleccionadas de la transformación se creó un modelo lineal y se evaluaron los resultado de las predicciones, los cuales se muestran a continuación.

set	RMSE	R^2
Test	2.1572	0.5517
Train	2.2680	0.5042

PLS

se ajustó el modelo PLS con diferente cantidad de variables, removiendolas iterativamente y evaluando el rendimiento de predicción en cada caso. Se observó que el cambio significativo en la capacidad de predicción se producía al ajustar con solo dos variables, por lo que se optó por utilizar 3 componentes.

Mediante el análisis gráfico, se pudo observar que las primeras 4 componentes mantenian una relación lineal con la variable dependiente.

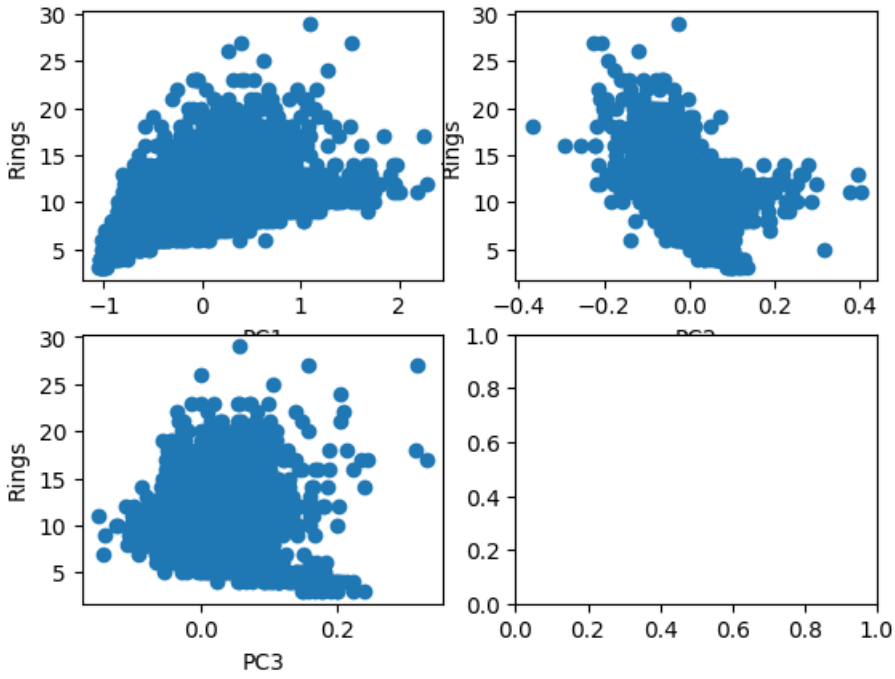

Figure 7: 3comp

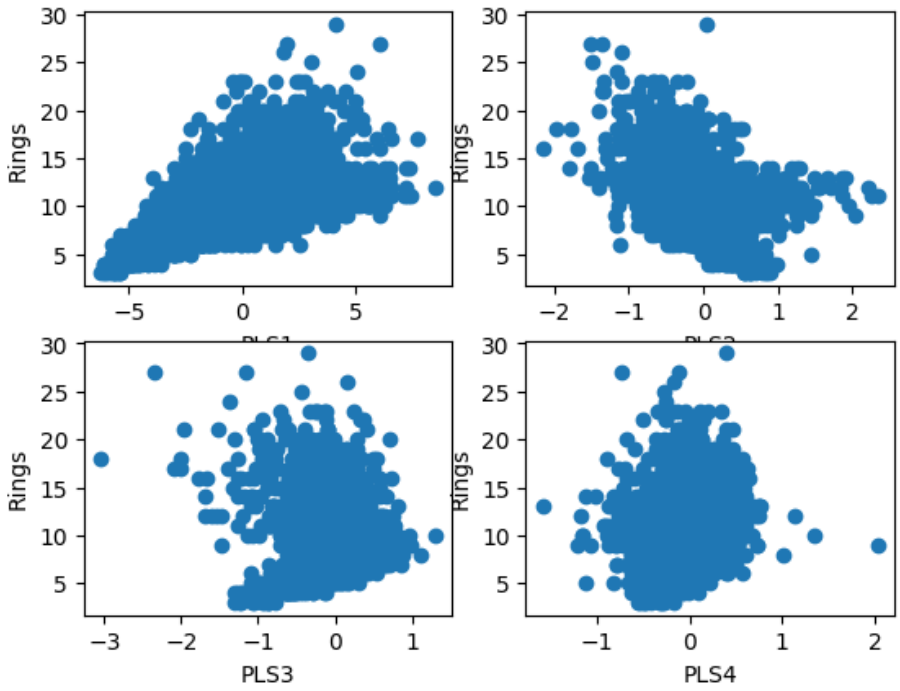

Figure 8: pls_comp

De el ajuste con tres componentes se obtuvo el siguiente puntaje.

set	RMSE	R^2
Test	2.1712	0.5459
Train	2.2559	0.5094

Concluciones

Se puede concluir que el modelo 'benchmark' fue el que mejor desempeño tuvo en términos de precisión, aunque utilizó el mayor número de variables. El modelo PCR presentó una precisión ligeramente menor, posiblemente debido a la reducción de variables. Por otro lado, el modelo PSL obtuvo resultados muy similares a pesar de usar el mismo número de componentes. Esta discrepancia con PCR sugiere que las variables con mayor varianza tienen mayor poder predictivo, aunque esto no siempre es cierto. Por último, la selección de variables basada en correlación obtuvo las métricas más bajas. La estrecha diferencia entre las métricas puede atribuirse a la alta correlación entre todas las variables independientes en la base de datos de abalón.