Análisis y Reporte sobre el desempeño del modelo

Enrique Santos Fraire

A01705746

Introducción

A través de un modelo de regresión lineal, se espera poder predecir el área de un conjunto de datos de pasas [1]. Dicho dataset cuenta con 7 características morfológicas y su especie, las cuales son:

- Area
- Perimeter
- MajorAxisLength
- MinorAxisLength
- Eccentricity
- ConvexArea
- Extent

La especie o class del conjunto de datos no será necesaria para el modelo. Se utilizarán las variables anteriores para estimar el área.

Modelo inicial

Para realizar el modelo del análisis se empleó la función LinearRegression [2] del framework SciKit-Learn [3], utilizando como parámetros:

- fit_intercept = True

Esto para calcular la intersección del modelo y ayude en la predicción de los datos.

Durante este modelo inicial se optó por utilizar la Eccentricity y el Perimeter como variables las variables independientes para entrenar el modelo.

Con el fin de analizar de una mejor manera el comportamiento de los datos y del modelo se separaron los datos en un conjunto de train para entrenar el modelo, y otro de test para evaluar el “fitting” del mismo.

Nuestros primeros resultados arrojaron lo siguiente:

```
r2 de train:  0.9302210104824834
r2 de test:  0.9481185406143141
Cross validation:  0.9288839922985004
```

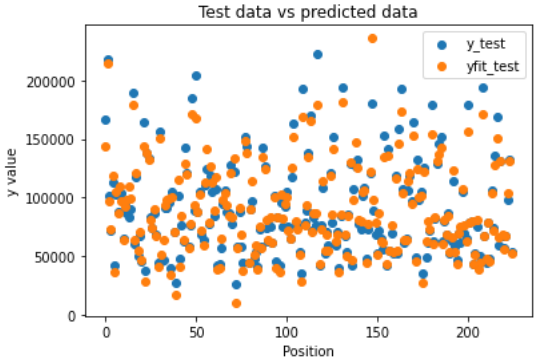

Fig 1: Gráfica de datos reales vs datos predichos (modelo inicial)

Como se puede ver, tenemos un buen modelo inicial, con un 93% de accurancy en los datos de entrenamiento, y un 94% en los datos de prueba, además de un 92% en la cross validation, lo que ya nos sitúa en un modelo “fit”, que puede a su vez ser corroborado en la figura 1, donde se comparan los datos reales de la prueba con los predichos, apreciándose un bias bajo dada la concentración de los puntos, aunque con una varianza ligeramente alta dada la dispersión.

Mejora del modelo

Si bien ya contamos con un modelo que puede considerarse “fit”, aún tenemos alternativas por explorar.

A pesar de la sencillez de la regresión lineal, aún hay capacidades de mejora por parte del uso de datos, como lo son los parámetros utilizados. Para ello, se buscó la correlación de las variables del dataset con nuestra variable dependiente, el área. Dándonos como resultado la siguiente tabla:

```
Area              1.000000
ConvexArea        0.995920
Perimeter         0.961352
MajorAxisLength   0.932774
MinorAxisLength   0.906650
Eccentricity      0.336107
Extent           -0.013499
Name: Area, dtype: float64
```

Fig 2: Tabla de correlación del Área

Se puede apreciar que el parámetro de Perimeter está fuertemente correlacionado con el área, mientras que la Eccentricity podía incluso ser dejada de lado. Es por eso que para nuestro nuevo modelo optaremos por utilizar como parámetros aquellas variables que tengan una correlación por encima del 90%, quedando descartadas las variables de Eccentricity y Extent.

Resultados

Una vez que se volvió a correr el modelo con estos nuevos parámetros se obtuvieron los siguientes resultados:

```
r2 de train:  0.996978646077254
r2 de test:  0.9977589223544987
Cross validation:  0.9966044232927187
```

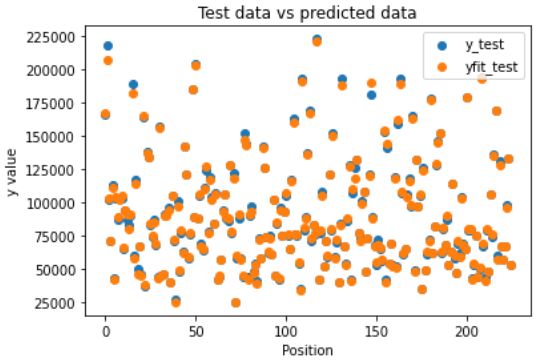

Fig 3: Gráfica de datos reales vs datos predichos (modelo mejorado)

Con estos nuevos parámetros ahora contamos con un 99% de accurancy tanto en los datos de entrenamiento, como en los de prueba, así como en la cross validation, manteniéndose como un modelo “fit”.

Por parte de la gráfica, nuevamente se puede apreciar un bias bajo y, esta vez, una varianza igualmente baja con una dispersión mucho menor que en el modelo inicial, confirmando así nuevamente la accurancy del modelo.

Conclusión

Se logró la predicción del área de las pasas con un 99% de precisión a través de un modelo de regresión lineal utilizando como parámetros aquellas variables con las que tenga mayor correlación con el área. Teniendo una mejoría del 7% en la cross validation en comparación al modelo inicial.

Referencias

[1] CINAR I., KOKLU M. and TASDEMIR S., (2020), Classification of Raisin Grains Using Machine Vision and Artificial Intelligence Methods. Gazi Journal of Engineering Sciences, vol. 6, no. 3, pp. 200-209, December, 2020. DOI: [https://archive.ics.uci.edu/ml/datasets/Raisin+Dataset]

[2] SciKit-Learn (s.f.) LinearRegression. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[3] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.