Tecnológico de Monterrey

Campus Querétaro

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el desempeño del modelo. (Portafolio Análisis)

Profesores

Benjamín Valdés Aguirre

Presenta

Cristian Rogelio Espinosa Díaz A01702752

Fecha:

10 de Septiembre de 2023

A través de un modelo de regresión lineal, se tiene el objetivo de realizar la predicción de la cantidad de calorías encontradas en alimentos de comida rápida de un conjunto de datos con información nutrimental de diversas cadenas de empresas dedicadas a la comida rápida, las cuáles son McDonald's, Burger King, Wendy's, Kentucky Fried Chicken, Taco Bell, Pizza Hut.

Dicho conjunto de datos posee la siguiente información sobre cada platillo de las empresas previamente mencionadas.

- Company: El restaurante donde se localiza el platillo
- Item: Nombre oficial del platillo
- Calories: Cantidad de calorías provenientes de todas las fuentes (unidad de medición "cal")
- Calories fromFat: Cantidad de calorías provenientes de grasa(unidad de medición "cal")
- Total Fat(g): Cantidad de grasas totales(unidad de medición "gramos")
- Saturated Fat(g): Cantidad de grasas saturadas(unidad de medición "gramos")
- Trans Fat(g): Cantidad de grasas trans(unidad de medición "gramos")
- Cholesterol(mg): Cantidad de colesterol(unidad de medición "miligramos")
- Sodium (mg): Cantidad de sodio (unidad de medición "gramos")
- Carbs(g): Cantidad de carbohidratos (unidad de medición "gramos")

Datos como el nombre de la compañía o el nombre del platillo no serán necesarios para el modelo debido a que se desea crear un modelo aplicable para un contexto general. Es decir, no se hará enfoque individual a cada restaurante.

Para la creación del modelo se empleó la función LinearRegression del framework SciKit-Learn utilizando el parámetro fit_intercept = True con el objetivo de calcular la intersección del modelo para hacer que la predicción de los datos sea más eficiente.

Para determinar qué variables utilizar se realizó una matriz de correlación para saber qué variables poseían la relación lineal más fuerte y significativa con nuestra variable a predecir.

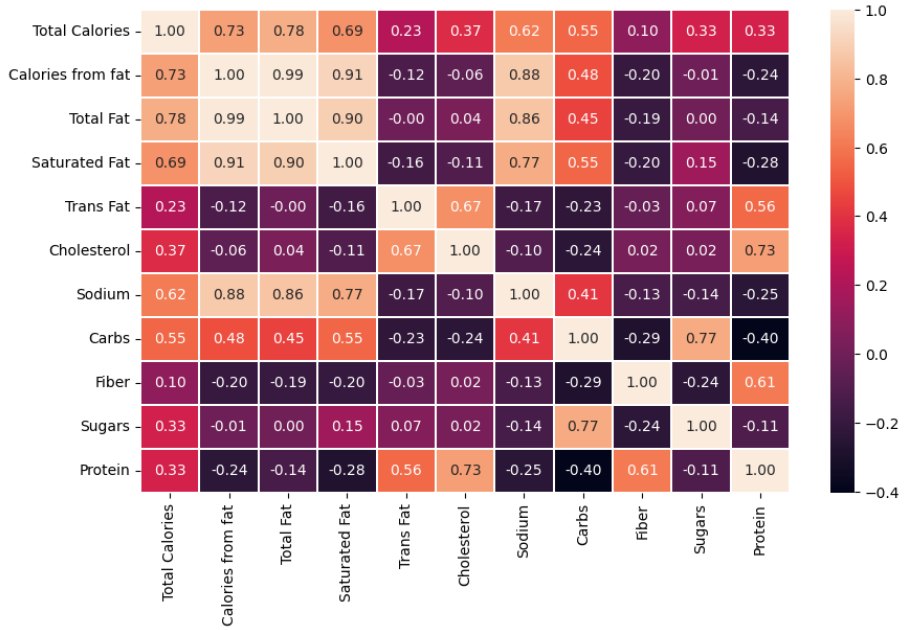

Fig 1: Matriz de correlación sobre la información nutrimental del dataset

Se optó por utilizar a Total Fat y Saturated Fat como variables independientes para entrenar el modelo al ser las que tenían la correlación positiva más alta.

- **Separación y evaluación del modelo con un conjunto de prueba y un conjunto de validación (Train/Test/Validation).**

Para tener un mejor acercamiento para analizar el comportamiento de los datos y del modelo, se dividió el set de datos en un conjunto de entrenamiento (train) y otro para las pruebas (test) para la evaluación del "fitting" de este modelo.

El 80% de los datos serán usados en la fase de train y el restante 20% será para la fase de test.

```
# Dividimos los datos en 80% para entrenar el modelo y 20% para el testeo del modelo
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)
```

A su vez, se utilizó la técnica de Cross-Validation con el objetivo de evaluar qué tan bien funciona el modelo. El objetivo es simular cómo el modelo se desempeñaría en datos nuevos y no vistos. En lugar de usar solo un conjunto de datos de entrenamiento y prueba, se divide el conjunto de datos en varias partes y se entrena y evalúa el modelo varias veces, de manera que cada parte se utiliza tanto para entrenamiento como para evaluación en diferentes momentos.

```
# Cross Validation del modelo
Cross_val = abs(cross_val_score(LinearRegression(), X_train, Y_train, cv=5, scoring = "r2").mean())
print ("Cross validation: ", Cross_val)
```

- **Desempeño del modelo.**

El error absoluto medio (MAE), el error cuadrático medio (MSE) y R-cuadrada serán las métricas utilizadas para evaluar el rendimiento de modelos de regresión.

El Error Absoluto Medio (Mean Absolute Error) mide cuán cerca están las predicciones de un modelo de regresión de los valores reales. Cuanto menor sea el MAE, mejor será el ajuste del modelo a los datos, ya que significa que las predicciones son más cercanas a los valores reales en términos de magnitud.

El Error Cuadrático Medio (Mean Squared Error) mide la magnitud promedio de los errores al cuadrado entre las predicciones de un modelo de regresión y los valores reales. Cuanto menor sea el MSE, mejor será el ajuste del modelo a los datos, ya que significa que las predicciones son más cercanas a los valores reales en términos de magnitud, y los errores son en su mayoría pequeños.

No obstante, esta métrica tiende a penalizar más fuertemente los errores grandes, lo que significa que valores atípicos pueden tener un impacto significativo en los resultados de dicha métrica.

R-cuadrado es una medida de cuánta variabilidad en la variable dependiente (objetivo) es explicada por el modelo de regresión en relación con la variabilidad total. Toma valores entre 0 y 1, donde 1 significa que el modelo explica toda la variabilidad y 0 significa que no explica nada.

Los resultados obtenidos en la ejecución de las diversas fases fueron las siguientes:

- Coeficiente para Total Fat: 7.3811538933865775
- Coeficiente para Saturated Fat: 4.534903975125634
- Intersección: -6.116855008810589

Fase de train

- MSE de train: 328.6665036220902
- MAE de train: 11.04768174233487
- R-squared de train: 0.9759148509692721

Fase de test

- MSE de test: 244.94319465563845
- MAE de test: 10.818165288126636
- R-squared de test: 0.9771518845014623

Fase de validación

- Cross validation: 0.9754729833174322

Análisis de los resultados

Un MSE y MAE bajos indican que el modelo hace predicciones cercanas a los valores reales. En este caso, tanto el MSE como el MAE son relativamente bajos en los conjuntos de entrenamiento y prueba, lo que indica un buen ajuste (fit) del modelo a los datos.

Tanto en el conjunto de entrenamiento como en el conjunto de prueba, el R-cuadrado es cercano a 1, lo que sugiere que el modelo explica la mayoría de la variabilidad en la variable objetivo.

El valor de validación cruzada también es alto (0.975) y está cerca de los valores de R-cuadrado en los conjuntos de entrenamiento (0.9759) y prueba (0.9771). Esto indica que el modelo generaliza bien a datos no vistos y no está sobre ajustado (overfit) a los datos de entrenamiento.

Un bajo MSE, MAE y alto R-cuadrado en el conjunto de prueba sugieren que el modelo tiene un bajo sesgo (bias) y una baja varianza. El buen rendimiento en la validación cruzada respalda la generalización del modelo.

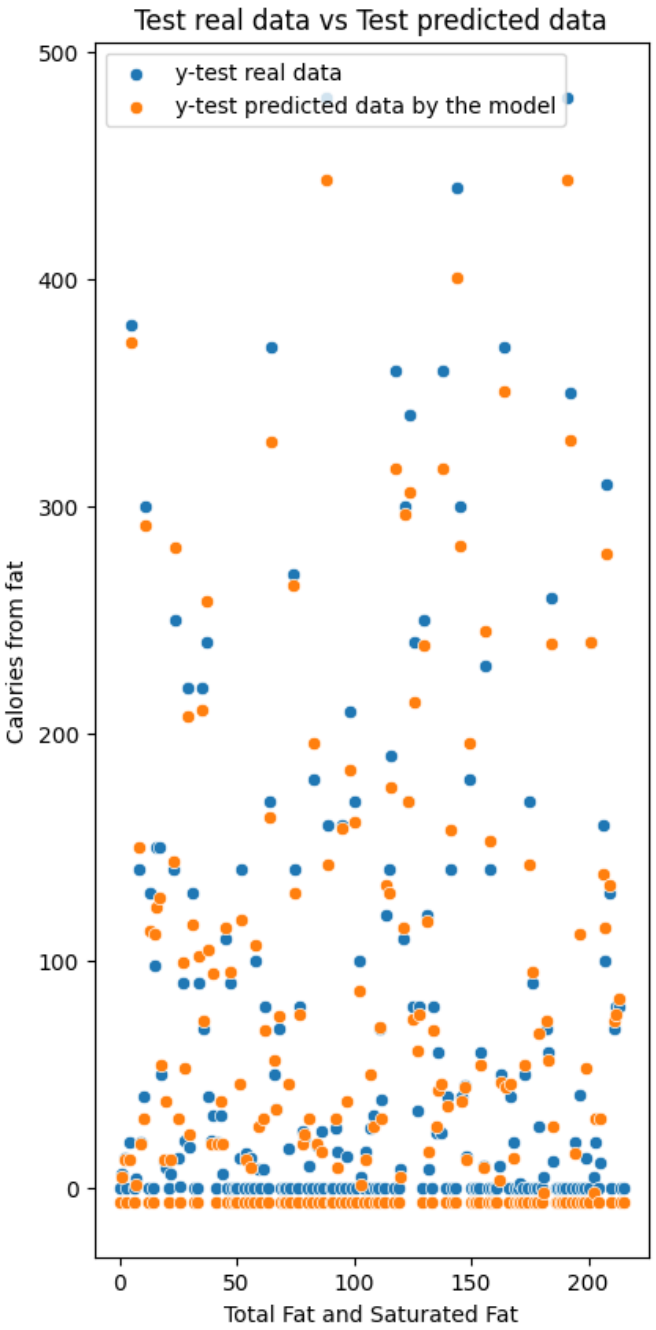

Fig 2: Gráfica de datos reales vs datos predecidos por el modelo (fase de test)

Mediante la Fig 2, donde se comparan los datos reales de test contra los que se predijeron por el modelo, podemos corroborar que el modelo se encuentra en la categoría de "bien ajustado" (fit) en lugar de "underfit" o "overfit". A su vez, se puede comprobar que hay un bias bajo dado la concentración de los puntos. No obstante, el gráfico también nos indica que hay una ligera varianza dada la dispersión de los puntos en la parte superior.

- **Mejora del modelo**

A pesar de que las métricas MAE, MSE y R-cuadrado nos muestran que el modelo tiene un buen equilibrio entre sesgo y varianza y tienen un nivel de ajuste deseado (fit). Hay un factor que se puede explorar y tener una mejora en el modelo, el cual es la varianza.

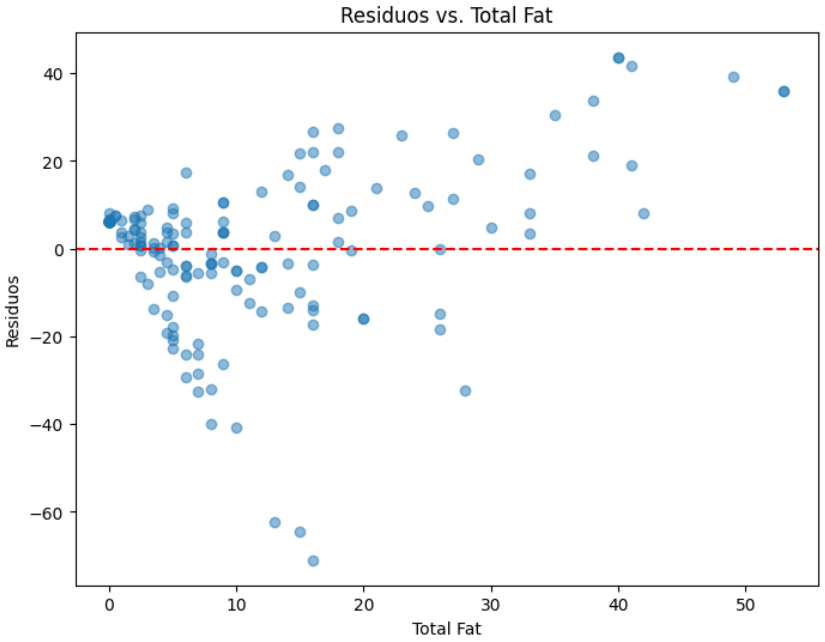

Fig 3: Gráfica de residuos de Total fat con datos de test

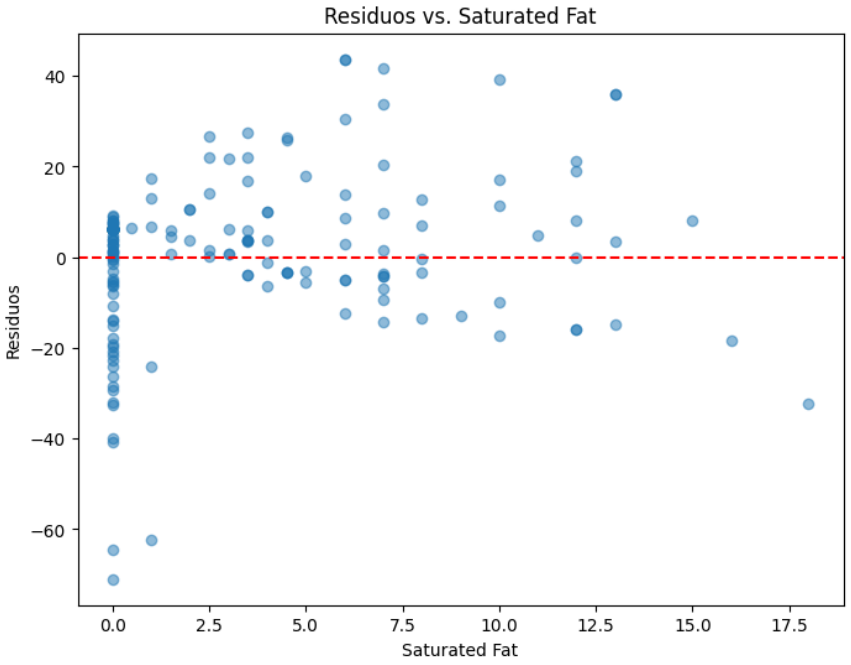

Fig 4: Gráfica de residuos de Saturated fat con datos de test

Al realizar la gráfica de residuos con los datos del conjunto de test se puede observar un patrón anómalo en “Saturated fat”. Esto nos está indicando una alta variabilidad en las predicciones del modelo. En otras palabras, si los residuos están dispersos de manera caótica y desigual, esto sugiere que el modelo tiene dificultades para predecir con precisión los valores observados y que la varianza de las predicciones puede ser alta.

Por otro lado, la dispersión en los residuos de "Total Fat" podría indicar que la relación entre esta variable y la variable objetivo (Calories from fat) es más variable y menos predecible por el modelo.

Por lo que se procederá a realizar la técnica de regularización Lasso (Least Absolute Shrinkage and Selection Operator) es utilizada para mejorar la capacidad de los modelos de regresión al controlar la complejidad del modelo y seleccionar automáticamente las características más importantes.

Lasso introduce una penalización adicional en la función de pérdida utilizada para ajustar la regresión lineal. Esta penalización se basa en la suma de los valores absolutos de los coeficientes de las características predictoras.

Permite encontrar un equilibrio entre sesgo y varianza en el modelo. A medida que aumenta el valor de alpha, se reduce la varianza (lo que evita el sobreajuste) a costa de un ligero aumento en el sesgo (lo que reduce la capacidad del modelo para ajustarse a los datos de entrenamiento). En otras palabras, Lasso ayuda a prevenir el sobreajuste al simplificar el modelo.

Modelo Inicial:

Fase de train

- MSE de train: 328.6665036220902
- MAE de train: 11.04768174233487
- R-squared de train: 0.9759148509692721

Fase de test

- MSE de test: 244.94319465563845
- MAE de test: 10.818165288126636
- R-squared de test: 0.9771518845014623

Fase de validación

- Cross validation: 0.9754729833174322

Modelo con Lasso:

Fase de train

- MSE de train lasso: 268.10115665322377
- MAE de train lasso: 9.901322461997028

- R-squared de train lasso: 0.9803531657709539

Fase de test

- MSE de test lasso: 226.2342310953585
- MAE de test lasso: 9.568753514243614
- R-squared de test lasso: 0.9788970424385267

Fase de validación

- Cross validation lasso: 0.9805269969491178

Comparando las métricas podemos realizar las siguientes afirmaciones:

- Reducción del MSE y MAE: Tanto el MSE (Error Cuadrático Medio) como el MAE (Error Absoluto Medio) son más bajos en el modelo Lasso en comparación con el modelo inicial, tanto en el conjunto de entrenamiento como en el de prueba. Esto indica que el modelo Lasso hace predicciones más precisas y tiene un mejor ajuste a los datos.

- Mejora en el R-cuadrado: Un R-cuadrado más elevado sugiere que el modelo Lasso es capaz de explicar una porción significativamente mayor de la variabilidad presente en la variable objetivo. En otras palabras, el modelo Lasso logra un impresionante R-cuadrado de 0.98, lo que indica que aproximadamente el 98% de la variabilidad en la variable que estamos tratando de predecir ha sido efectivamente explicada por el modelo de regresión y las variables predictoras que se han incorporado en él.

- Validación Cruzada: La validación cruzada también muestra una mejora en el modelo Lasso en comparación con el modelo inicial. Esto indica que el modelo Lasso es más robusto y generaliza mejor a datos no vistos.

Bias/Sesgo del modelo Lasso:

En el modelo Lasso, el MAE de entrenamiento es de 9.90, mientras que en el modelo inicial era de 11.05. Esto indica una reducción en el error absoluto medio. Una reducción similar se puede observar con el MAE de prueba ya que en el modelo inicial es de 10.82 mientras que en Lasso es de 9.57.

Lo que indica que el modelo Lasso tiene un sesgo ligeramente menor en comparación con el modelo inicial ya que el MAE es más bajo en todas las fases..

Varianza del modelo Lasso:

El MSE de entrenamiento del modelo Lasso es de 268.10, en comparación con 328.67 en el modelo inicial. El MSE de prueba del modelo Lasso es de 226.23, en comparación con 244.94 en el modelo inicial. Aunque el MSE del modelo Lasso es ligeramente mayor en el conjunto de prueba, sigue siendo un valor bastante bajo en comparación con el modelo inicial.

En general, el modelo Lasso parece tener una varianza igual o ligeramente menor en comparación con el modelo inicial, lo que sugiere que ambos modelos tienen una capacidad similar para lidiar con la variabilidad de los datos.

Nivel de ajuste del modelo Lasso:

El R-cuadrado de entrenamiento del modelo Lasso es de 0.980, mientras que en el modelo inicial era de 0.976. El R-cuadrado de prueba del modelo Lasso es de 0.979, en comparación con 0.977 en el modelo inicial.

En resumen, el nivel de ajuste (fit) del modelo Lasso es igual o ligeramente mejor que el del modelo inicial, ya que el R-cuadrado es ligeramente más alto en ambas fases.

Validación Cruzada del modelo Lasso:

La validación cruzada también muestra una mejora en el modelo Lasso en comparación con el modelo inicial. Esto indica que el modelo Lasso es más robusto y generaliza mejor a datos no vistos.

En general, el modelo Lasso parece mejorar en términos de sesgo y ajuste en comparación con el modelo inicial, manteniendo una varianza similar. Sin embargo, es importante destacar que las diferencias son relativamente pequeñas, lo que indica que ambos modelos son bastante buenos en términos de rendimiento.

Referencias:

Fernández Casal, R., Costa Bouzas, J., & Oviedo de la Fuente, M. (Eds.). (2021). Aprendizaje Estadístico. https://rubenfcasal.github.io/aprendizaje_estadistico/shrinkage.html#