Proyecto de ML para incrementar el gasto anual de clientes en la tienda.

Evaluación del Modelo.

De la **'validación cruzada'** obtenemos:

MSE para cada conjunto: 128.3, 107.8, 137.8, 105.0, 91.2, 75.5, 149.6, 130.1, 91.7 y 103.8.
Media para todos los conjuntos: 112.1.
Desviación estándar para todos los conjuntos: 22,3
R^2 Score de cada conjunto: 0.983, 0.981, 0.980, 0.974, 0.981, 0.990, 0.980, 0.979, 0.984, 0.980.
R^2 Score Promedio: 0.981

Se observa que los valores para el error cuadrático son relativamente pequeños y muestran poca dispersión.

Como se ve, R^2 para todos los conjuntos de datos es muy próximo a 1.0. Un valor próximo a 1 significa que los errores son pequeños y que, en efecto, el modelo de una recta se ajusta a los datos. En particular, 'Yearly Amount spent' crece con 'Lenght of Membership', 'Time on App' y 'Avg. Session Lenght' según una linea recta.

De la **'validación completa'** obtenemos:

MSE: 162.5
R^2 Score: 0.976
Coeficientes de regresión lineal: 61.43, 39.28, 24.77.
Intercept: 500.38.

Vuelve a comprobarse que el error MSE es pequeño y el coeficiente R^2 próximo a 1.0, el modelo lineal se ajusta correctamente.

Los coeficientes muestran como varía la variable objetivo según cada caracterísca. Por ejemplo, si 'Length of Membership' se incrementa en 1 mes, 'Yearly Amount spent' aumentaría en 61.43€, asumiendo que 'Time on App' y 'Avg. Session Lenght' fueran 0. Si los coeficientes son positivos 'Yearly Amount spent' crece si las características crecen. Si fueran negativos, 'Yearly Amount spent' disminuría al aumentar las características. 'Intercept' es el valor de 'Yearly Amount spent' cuando las otras características son 0.

La gráfica de los valores reales respecto de la predicción recalca la pequeña desviación de los errores y por tanto que el modelo está bien ajustado. Sin embargo, se observan valores alejados de la recta ideal y, por otro lado, un valor de MSE igual a 162 en la validación completa resulta algo alto. Un MSE alto puede indicar que la diferencia entre el valor real y el predicho, el error, es demasiado alta. Es una métrica, además, que penaliza los errores altos. Para mejorar dichos datos se podría intentar aplicar diferentes modelos, como regresión polinómica, SVR o 'decision tree regression'. También se podría comprobar si hay características que estén introduciendo ruido y eliminarlas. Por último, se podrían ajustar hiperparámetros como la fuerza de regularización (Alpha), la tasa de aprendizaje o el número de iteraciones.

Como conclusión el modelo muestra una correlación lineal entre 'Lenght of Membership', 'Time on App' y 'Avg. Session Lenght' con 'Yearly Amount spent'. Por tanto, con un tiempo mayor como usuario registrado en la aplicación WEB y en la aplicación móvil, es más probable que el gasto aumente. Asímismo, resulta más probable que el gasto aumente con una mayor interacción del cliente con la WEB y la aplicación de móvil.

Lluís Sala - 6/10/2024