

PRÁCTICA PEC2 ENTREGABLE: PREDICCIÓN DINÁMICA MEDIANTE MODELOS LINEALES, KNN Y ÁRBOLES





Introducción al problema

Nos encontramos en un problema que tiene que ver con predicciones de cotizaciones de S&P500, lo primero de todo para obtener los datos del 1 de enero a 2021 hasta 28 de febrero de 2023 en yahoo finance no nos daba la opción de descargarlos entonces conseguimos los datos en investing.com.

Seguidamente los datos nos lo daban ordenados por fecha, lo que tuvimos que invertir dicho dataframe y usar los datos de la columna Último ya que son los datos de cierre de cotización de ese día, para poder calcular las rentabilidades y poder empezar a abordar el problema.

Hemos creado dos matrices una de valores independientes (X) y otra de valores dependientes (Y), se debe a que vamos a usar un aprendizaje supervisado, seguidamente en nuestro código se puede apreciar como hemos realizado cada modelo obteniendo sus métricas correspondientes.

1. ¿De cuantos parámetros consta el modelo?

$$Y = a + b_1r_{t1} + b_2r_{t2} + b_3r_{t3} + b_4r_{t4} + b_5r_{t5}$$

Cómo vamos a calcular la rentabilidad de los 5 días anteriores. Podemos apreciar en la fórmula que consta de 5 parámetros que dependen de **bn** (pesos) y la otra variable continua que es **a** que está multiplicado por 1 (ones).

En total hay 6 parámetros para el modelo lineal.



2. ¿Los parámetros correspondientes al tercer y cuarto regresor son iguales?

No pueden ser iguales ya que si lo fueran estaríamos dando por hecho que los valores de cierres han de ser los mismos para poder tener las mismas rentabilidades .

Puede ocurrir con una probabilidad baja pero para decir que son iguales debería de ocurrir de manera constante lo que es imposible ya que no se puede predecir cotizaciones con un 100% de exactitud.

Las cotizaciones dependen de diversos factores lo que dificulta mucho su predicción.

3. ¿Las tres métricas, proporcionan los mismos resultados (el mismo valor)?

No, porque, aunque las tres miden el error entre la predicción y los valores reales, tienen un enfoque diferente para hacerlo. El mse magnifica los errores más grandes ya que lo eleva al cuadrado. El mae, ya que opera en valores absolutos, le da la misma importancia a todos los valores y penaliza los errores de manera uniforme. El mape calcula el porcentaje promedio de error en las predicciones.

Metricas del modelo:

Exactitud Modelo Lineal: 0.0026

MAE: 0.010916

MSE: 0.000195

MAPE: 0.982741

$$Y = [5.38040186e-05] + Rt1 * [-0.00209109] + Rt2 * [-0.0206224] + Rt3 * [-0.02825086] + Rt4 * [0.01119571] + Rt5 * [-0.03612535]$$

4. ¿Suponga que comparamos las predicciones frente a un paseo aleatorio $rt=0$ empleando el mape, las predicciones del modelo lineal son mejores?

Las predicciones del modelo lineal, dadas las métricas utilizadas, son mejores que las del paseo aleatorio con $rt=0$.

El MAE del modelo lineal es especialmente menor que el del paseo aleatorio, por lo que las predicciones son más precisas en términos de la magnitud del error promedio.


El MSE del modelo lineal es menor que el del paseo aleatorio, esto significa que el modelo lineal tiene una mejor capacidad para capturar las variaciones del rendimiento.

También el MAPE es mucho más pequeño que en el paseo aleatorio, lo que indica que las predicciones del modelo lineal son más precisas en términos de porcentaje de error promedio.

```
-----MAE-----  
Paseo Aleatorio : 0.013844  
Lineal: 0.010916
```

```
-----MSE-----  
Paseo Aleatorio: 0.000320  
Lineal: 0.000195
```

```
-----MAPE-----  
Paseo Aleatorio: 3.016012  
Lineal: 0.982741
```



5. Encuentre, sobre el conjunto de entrenamiento y empleando validación cruzada (10 conjuntos) el mejor modelo de vecinos próximos (algoritmo knn) empleando entre 1 y 15 vecinos. Con el modelo óptimo calcule el ecm sobre el conjunto de test y compárelo con un paseo aleatorio. ¿Es mejor o peor que un paseo aleatorio?

Dados los resultados, es mejor el modelo de vecinos próximos que un paseo aleatorio porque tiene un valor de ecm menor (0,000121 para vecinos próximos y 0,000279 para paseo aleatorio). Esto indica que el modelo de vecinos próximos tiene una mejor capacidad para predecir los precios futuros.

Además, obtenemos que el número óptimo de vecinos son 13, ya que es probable que al probar valores más bajos de k , el modelo no haya sido capaz de capturar patrones más complejos en los datos y, por lo tanto, haya obtenido predicciones menos precisas.

5. Encuentre, sobre el conjunto de entrenamiento y empleando validación cruzada (10 conjuntos) el mejor modelo de vecinos próximos (algoritmo knn) empleando entre 1 y 15 vecinos. Con el modelo óptimo calcule el ecm sobre el conjunto de test y compárelo con un paseo aleatorio. ¿Es mejor o peor que un paseo aleatorio?

Pódemos observar en la gráfica que hemos hecho para representar el ECM (MSE) en el eje de ordenadas y N° de vecinos en el eje de abscisas.

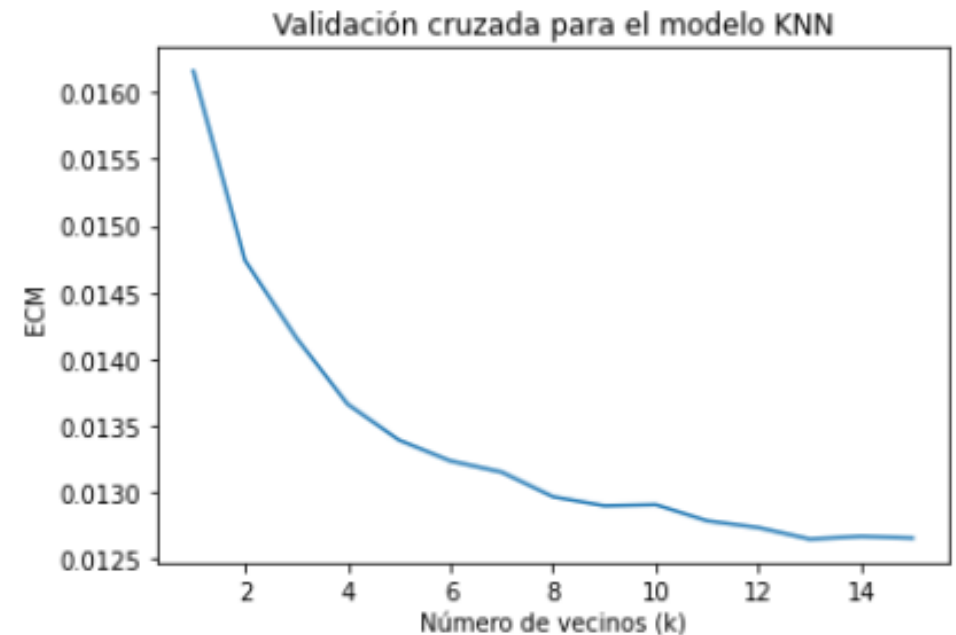
El ECM menor es el del KNN como hemos comentado anteriormente, gracias a la representación gráfica podemos visualizar mejor con qué número de vecinos se obtiene el menor error cuadrático medio.

Exactitud knn: 0.08

ECM en el conjunto de KNN: 0.000121

ECM en el conjunto de paseo Aleatorio: 0.000279

El número optimo de vecinos son: 13





6. Empleando el mismo procedimiento, métrica y benchmark con un árbol de regresión con una profundidad entre 1 y 10 niveles. ¿Es mejor o peor que un paseo aleatorio?

El ECM es una medida de la diferencia entre las predicciones del modelo y los valores reales y, por lo general, un valor más bajo de esta medida indica un mejor ajuste del modelo a los datos.

Cuanto menor sea el valor del ECM, mejor es la capacidad del árbol de regresión para ajustarse a los datos de entrenamiento y generalizar datos nuevos.

A medida que el árbol va teniendo más profundidad, el ecm va siendo mayor y se va igualando bastante con los valores de ecm del paseo aleatorio.

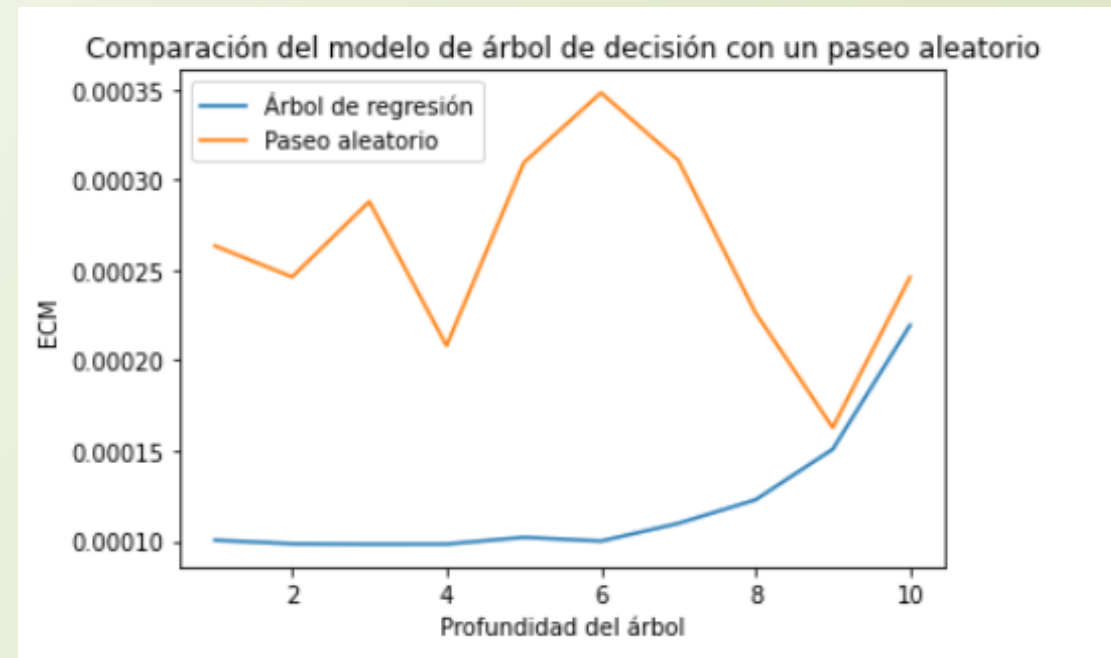
Esto es debido a que una profundidad más grande permite que el árbol capture más detalles en los datos de entrenamiento, pero también tiene el peligro de que pueda llevar a un sobre ajustamiento del modelo de los datos de entrenamiento.

6. Empleando el mismo procedimiento, métrica y benchmark con un árbol de regresión con una profundidad entre 1 y 10 niveles. ¿Es mejor o peor que un paseo aleatorio?

En este caso, hemos evaluado un modelo de árbol de decisión con diferentes profundidades, de 1 a 10.

Teniendo en cuenta los resultados que hemos obtenido, podemos observar que el árbol de decisión con profundidad 1 tiene un ECM mucho más bajo que el paseo aleatorio. Pero a medida que aumenta la profundidad del árbol, el ECM en el conjunto de validación aumenta.

Esto indica que el modelo de árbol de decisión con profundidad 3 es el mejor ajuste para estos datos, ya que tiene el ECM más bajo en el conjunto de validación.



6. Empleando el mismo procedimiento, métrica y benchmark con un árbol de regresión con una profundidad entre 1 y 10 niveles. ¿Es mejor o peor que un paseo aleatorio?

Profundidad del árbol: 1
ECM en - Árbol: 0.000100
ECM en - Paseo aleatorio: 0.000263

Profundidad del árbol: 2
ECM en - Árbol: 0.000098
ECM en - Paseo aleatorio: 0.000246

Profundidad del árbol: 3
ECM en - Árbol: 0.000098
ECM en - Paseo aleatorio: 0.000288

Profundidad del árbol: 4
ECM en - Árbol: 0.000098
ECM en - Paseo aleatorio: 0.000208

Profundidad del árbol: 5
ECM en - Árbol: 0.000102
ECM en - Paseo aleatorio: 0.000309

Profundidad del árbol: 6
ECM en - Árbol: 0.000100
ECM en - Paseo aleatorio: 0.000348

Profundidad del árbol: 7
ECM en - Árbol: 0.000110
ECM en - Paseo aleatorio: 0.000311


Profundidad del árbol: 8
ECM en - Árbol: 0.000123
ECM en - Paseo aleatorio: 0.000226

Profundidad del árbol: 9
ECM en - Árbol: 0.000151
ECM en - Paseo aleatorio: 0.000163

Profundidad del árbol: 10
ECM en - Árbol: 0.000219
ECM en - Paseo aleatorio: 0.000246

Como conclusión, el modelo de árbol de regresión es mejor que el paseo aleatorio, en relación con los resultados de ECM, cuanto menor sea su profundidad.

Además, observamos que la exactitud o precisión del árbol tiene un valor de 0.50, lo que representa cómo de cerca están los valores predichos y los valores reales. Este valor indica que el modelo puede explicar el 50% de la variabilidad en los datos de entrenamiento. Además de que predice correctamente la mitad de las observaciones y, en la otra mitad, se aleja de los valores reales.




7. ¿La rentabilidad de cada uno de los modelos (lineal, knn y árboles) es positiva?

Hemos utilizado 5 métricas para valorar la rentabilidad de cada modelo: el coeficiente de determinación R^2 , el error absoluto medio (mae), y el error cuadrático medio (mse), la exactitud del modelo y el error porcentual absoluto medio (mape).

El coeficiente de determinación es una métrica que se utiliza para evaluar la calidad y rentabilidad de un modelo de regresión. Determina cuánta variación de la variable de respuesta es explicada por el modelo. R^2 siempre toma valores entre 0 y 1. Un resultado cercano a 0 indica que el modelo no explica nada de esta variación, mientras que un valor cercano a 1 indica que el modelo sí explica esta variabilidad.

Podemos obtener valores negativos del coeficiente de determinación R^2 cuando el modelo ajustado es tan malo que el modelo no explica la variación de la variable de respuesta y además, podríamos decir que la hace peor. Esto sucede cuando el modelo se ajusta a los datos de una forma muy pobre, lo que resulta una predicción peor que simplemente usar el valor medio de la variable de respuesta.



7. ¿La rentabilidad de cada uno de los modelos (lineal, knn y árboles) es positiva?


El error absoluto medio (mae), también es una medida de la calidad y rentabilidad de las predicciones realizadas por un modelo de regresión. Esta medida indica cuánto se espera que varíe la salida del modelo cuando se realiza una predicción para una entrada que no se ha producido antes.

Si el mae es bajo, indica que el modelo es capaz de realizar predicciones precisas y, por lo tanto, es rentable. Por lo contrario, si el mae es alto, significa que el modelo no es preciso a la hora de realizar predicciones ante valores reales no vistos anteriormente y, por lo tanto, no es rentable.

El error cuadrático medio (mse), también es una medida de calidad y rentabilidad de las predicciones de se realizan por un modelo de regresión e indica la variación de los resultados entre la predicción y una entrada que no se ha visto antes. Se calcula como el promedio de los cuadrados de las diferencias entre las predicciones del modelo y los valores reales.

En general, cuanto menor es el valor del mse, mejor es la calidad de las predicciones del modelo. Ya que en nuestro caso utilizamos modelos de regresión, podemos utilizar esta medida para evaluar la rentabilidad de nuestros modelos y cuantificar el error de predicción promedio del modelo.

En resumen, si el mse es bajo, indica que el modelo tiene un bajo error promedio en sus predicciones y, por lo tanto, puede ser rentable. Por otro lado, si el mse es alto, podría indicar que el error promedio de predicción es alto y, por lo tanto, sería recomendable cambiar el modelo por otro más rentable.



7. ¿La rentabilidad de cada uno de los modelos (lineal, knn y árboles) es positiva?

Cuanto menor sea el MAPE, mejor será la capacidad del modelo para predecir con precisión los valores futuros. Un MAPE bajo indica que el modelo de regresión tiene una buena capacidad de predicción y que las predicciones son precisas en términos porcentuales. Por otro lado, un mape alto indica que el modelo tiene una menor capacidad de predicción y que las predicciones están lejos de los valores reales.

La exactitud de un modelo de regresión indica cómo de bien el modelo podría predecir la rentabilidad futura de un valor no conocido. Un modelo preciso con alta capacidad de predicción es el que obtiene un valor de su exactitud próximo a 1, mientras que el modelo que obtiene una exactitud cercana a 0 no tiene una buena predicción.

7. ¿La rentabilidad de cada uno de los modelos (lineal, knn y árboles) es positiva?

-----MAE-----
Lineal: 0.007812
KNN: 0.005238
Árbol: 0.009022

-----MSE-----
Lineal: 0.000071
KNN: 0.000037
Árbol: 0.000090

-----MAPE-----
Lineal: 2.481425
KNN: 1.587772
Árbol: 2.923318

Exactitud modelo lineal: 1.0000

Exactitud modelo KNN: 0.0000

Exactitud modelo Arbol: 0.6737

-----Coeficientes Determinacion-----
Lineal: -6.831659
KNN: -3.018879
Árbol: -8.956819

Modelo de regresión lineal: la exactitud del modelo es 1, indica que las predicciones del modelo coinciden exactamente con los valores reales. Sin embargo, el R2 es negativo, lo que indica que el modelo no es adecuado para explicar la relación entre las predicciones y los valores reales y, por lo tanto, no es rentable. Así, podemos pensar que el modelo está sobreajustado a los datos de entrenamiento y no generaliza bien nuevos datos.

Además, el MAE y el MSE son cercanos a 0, indica que las predicciones difieren poco de los valores reales y que el modelo tiene buena capacidad para ajustarse a los datos. El MAPE es de 2,48 lo que significa que las predicciones difieren de los valores reales en un 2,48% y que el modelo debe tener una buena capacidad de predicción.

Modelo de vecinos próximos: la exactitud del modelo es 0, lo que indica que las predicciones del modelo no coinciden en absoluto con los valores reales. Además, el coeficiente de determinación es negativo, por lo que el modelo no es adecuado para explicar la variabilidad de respuesta y el modelo no es rentable.

Por otro lado, el MAE y el MSE son cercanos a 0, dando la misma respuesta que el modelo lineal. El MAPE es de 1.58, lo que significa que las predicciones difieren de los valores reales en un 1,58% y que el modelo debe tener una buena capacidad de predicción.

Modelo de árbol: la exactitud del modelo es 0.67, lo que indica que las predicciones tienen una precisión moderada en comparación con los valores reales y, además, el R2 es negativo, por lo que el modelo no parece rentable ya que no explica la variabilidad de respuesta.

El MAE y el MSE son muy cercanos a 0, dando buenos resultados en relación con cuánto difieren las predicciones de los valores reales y cuánta capacidad tiene para ajustarse a los datos de entrenamiento. El MAPE es de 2.923, lo que significa que en promedio las predicciones difieren en un 2.92% de los valores reales.

En conclusión, el modelo más rentable parece el modelo lineal, ya que, aunque tenga un coeficiente R2 negativo, también tiene los buenos resultados en el MAE, MSE y MAPE en comparación con los otros dos modelos y, además, una exactitud total. El modelo de vecinos próximos tiene los mejores resultados en el MAE, MSE y MAPE, pero muy malos resultados en el coeficiente de determinación y la exactitud.



8. Sugerir posibles mejoras a los procedimientos efectuados en 5. 6. y 7.

Mejoras procedimiento 5:

- Uso de la métricas recall o F1-score. Se puede usar recall para intentar minimizar los falsos negativos o F1-score si queremos encontrar un punto medio entre recall y el de precisión (usado actualmente).
- Aplicación de la validación cruzada estratificada o dejando uno fuera. Dejando uno fuera es útil en conjunto con datos pequeños ya que evalúa una instancia y las otras son de entrenamiento, es útil para conjuntos de datos pequeños. La estratificada, consiste en equilibrar los datos de entrenamiento y evaluación cuando se tienen conjuntos de datos desequilibrados (como es nuestro caso), ayudando a obtener una evaluación más precisa.
- Distintos valores de k. El objetivo sería observar cómo afecta al rendimiento otro rango de valores de k (actual 1-15), buscando un valor óptimo de k para este modelo.



8. Sugerir posibles mejoras a los procedimientos efectuados en 5. 6. y 7.

Mejoras procedimiento 6:

- Distintas profundidades del árbol. En este caso la profundidad se establece en el rango 1-10. Si se probasen profundidades más altas, puede que se encuentre una profundidad que proporcione mejor rendimiento que las ya probadas.
- Considerar otros parámetros. En este ejercicio se tiene en cuenta la profundidad pero podrían aplicarse otros parámetros como el número de muestras para dividir un nodo o incluso combinarse para que proporcione el mejor rendimiento.
- Otros métodos de comparación. Método de regresión lineal o aprendizaje profundo pueden aplicarse para evaluar su rendimiento con respecto a los dos ya aplicados.

Mejoras procedimiento 7:

- RMSE. Se podría tener en cuenta también esta medida para valorar el error de predicción de los modelos, cuanto menor es el valor, mejor es la precisión.
- Cambio de tamaño de ventana. Se puede tener en cuenta un tamaño de ventana mayor, donde se puede ver con más claridad aún que el modelo lineal es mejor ya que cuanto con mayor valor de tamaño de ventana, más aumenta dónde podría dejar de ser negativo.
- Aumentar el tamaño de la matriz. Al coger solo 8 días obtenemos unos datos de test que es posible que no sean muy representativos. Si cogiéramos más días aseguraríamos con mayor precisión la validez de los modelos