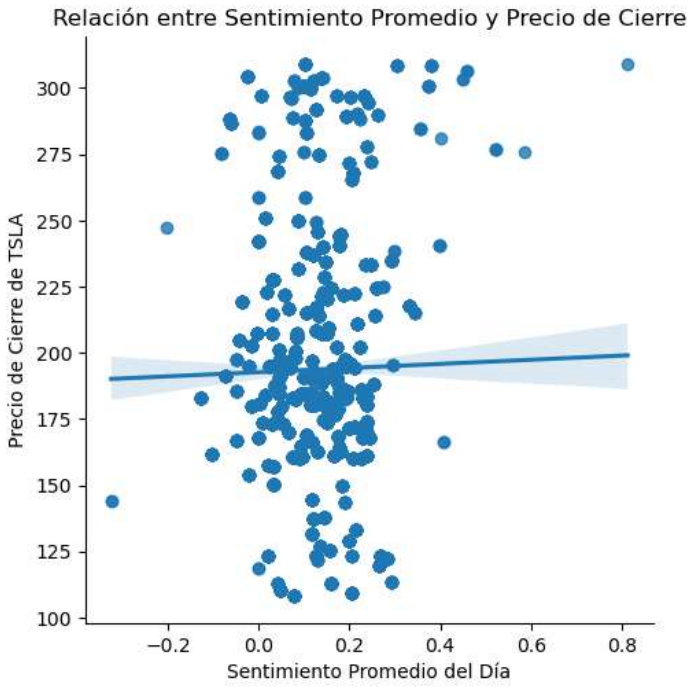
Juan Diego Sánchez

00344455

07/04/2025

Feature Engineering

En primer lugar, se busca una correlación entre el sentimiento promedio y el precio de cierre, para lo cual se agrupa el sentimiento por día y se compara contra el precio de cierre, se calcula la correlación obteniendo un 0,0159 lo cual se confirma con el siguiente grafico



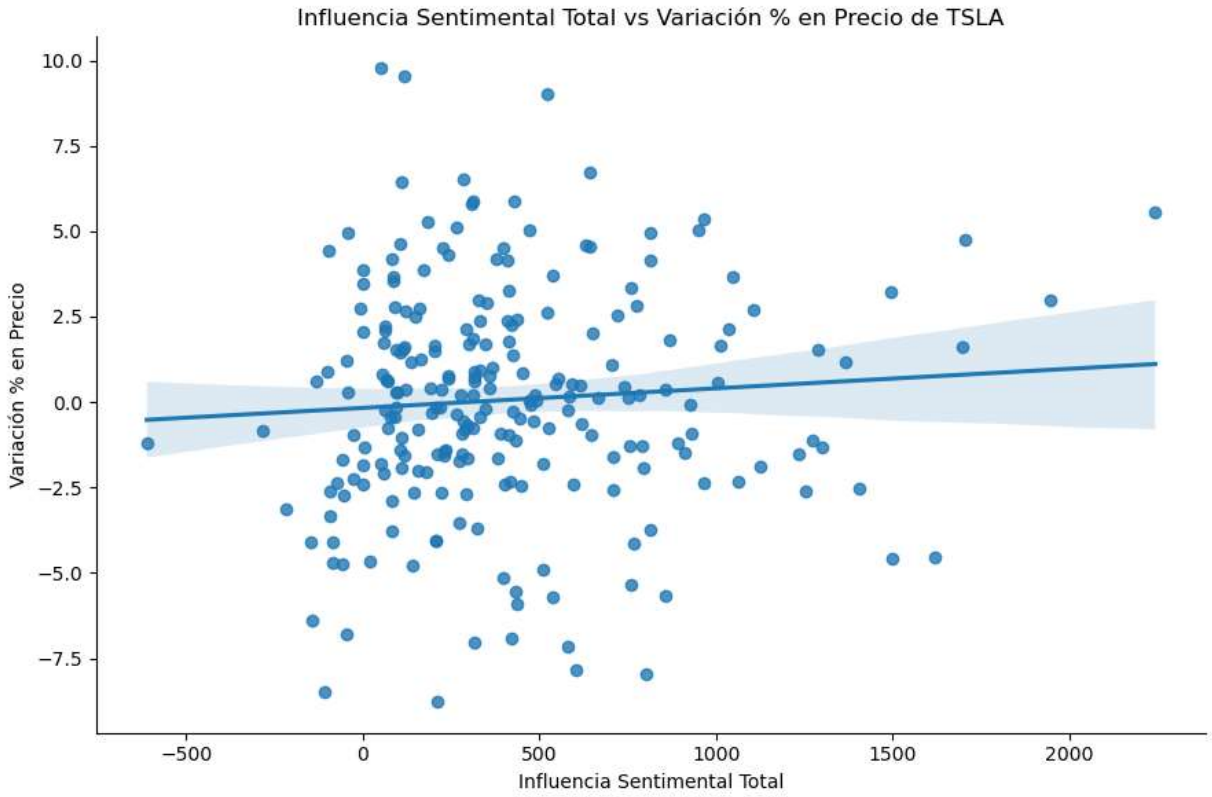
Posteriormente, se buscó validar la correlación, pero con el sentimiento promedio del día anterior, alcanzado una correlación de 0,0162.

Luego, al no alcanzar una correlación lo suficientemente significativa se procede a generar dos nuevas características que son la variación diaria del precio de la acción tanto en dólares como en porcentaje, eso se lleva a cabo calculando la resta del precio de cierre con el precio de apertura. Aquí se alcanza una correlación de 0,1532. Se prueba de igual manera la correlación contra 1 dia de desfase y con la media móvil de 3 días alcanzando 0,141 y 0,149 respectivamente.

A continuación se busca si la impacto en los días donde el sentimiento fue mas extremo produce un mayor impacto en la variación del precio para hacer esto se filtra solo los días con sentimiento promedio mayor a 0,5. En este escenario se alcanza una correlación muy alta (0,928) pero con un numero muy pequeño de muestras por lo que no podrían ser utilizadas para un modelo.

Posteriormente y bajo el criterio de ponderar el sentimiento con la amplificación pública se crea una característica adicional ponderando la cantidad de veces que fue retweeteado una publicación junto con el

sentimiento cuantificado. Se procede de igual manera con las características de seguidores, favoritos y finalmente se cuantifica la influencia total juntando todas las variables anteriores.

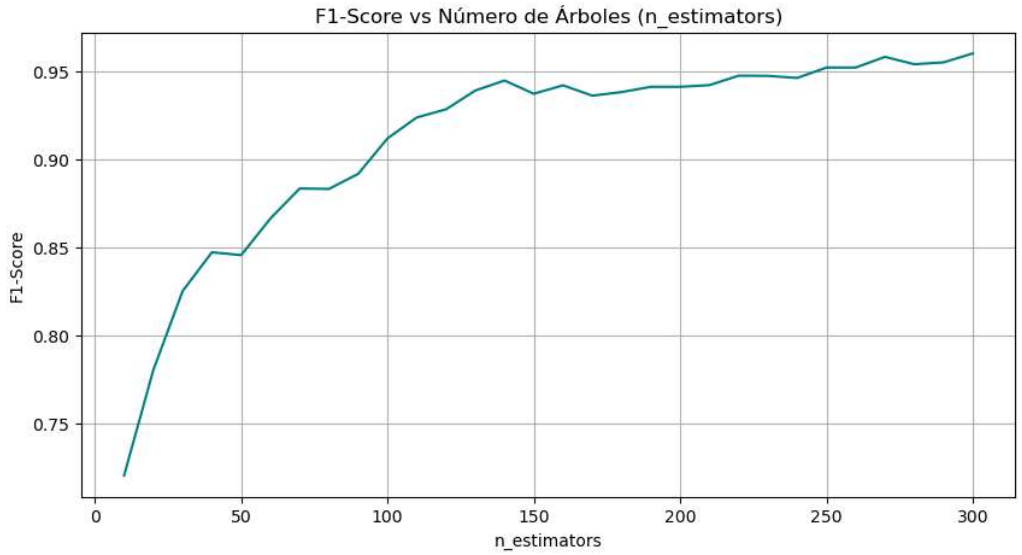


Modelos Evaluados

Se empieza el análisis con un random forest classifier alcanzando un f1 de 0,91 y un f1 de validación cruzada de 0,525. También se cuantifica la importancia de las características y se grafica concluyendo que las más importantes son el sentimiento promedio, sentimiento con lag de 1 día y sentimiento promedio de 3 días. Es decir, el modelo se esta enfocando en el sentimiento diario y la tendencia de sentimiento colectivo.

Luego, se procede a comparar la regresión lineal, voting classifer, bagging classifier y gradient boosting classifer alcanzando f1s de 0.573, 0.956, 0.971 y 0.977 respectivamente.

Finalmente se decide por el gradiente boosting classifier y se procede a ajustar los parámetros para lo cual se toma un enfoque gráfico mostrado a continuación y se decide por n_estimators igual a 160 y learning rate de 0.3:



Lecciones Aprendidas

- Se identifica un poder predictivo limitado del sentimiento, él cual logra mejorarse con métricas de amplificación como: retweets, favoritos y seguidores
- Modelos como Gradient Boosting ofrece buenos resultados, pero debe ser regularizado para evitar overfitting y memorización de patrones
- Es posible identificar correlaciones entre sentimiento y comportamiento del precio de la acción que sean estadísticamente relevantes