2023 - 2024

ACTIVIDAD GUIADA 1

Guía didáctica

Máster en Big Data y Data Science

05MBID – Estadística Avanzada

Nombre: Brahian Andrey Giraldo Alzate

Fecha: 08/30/2023

viu Universidad Internacional de Valencia

Contenido

1 Introducción, motivación y objetivo (Entendimiento del dominio)

La aplicación de técnicas de análisis de regresiones juega un papel fundamental en la búsqueda de patrones y tendencias en los datos financieros. Se realizará un análisis exhaustivo de datos financieros, utilizando herramientas estadísticas para modelar relaciones y construir estrategias de trading automatizadas.

El propósito de este trabajo es explorar cómo las técnicas de regresión pueden ser empleadas en el ámbito del trading algorítmico.

El análisis de los datos se llevará a cabo utilizando el lenguaje de programación R, una herramienta versátil que permite realizar cálculos estadísticos precisos y modelado avanzado. Se proporcionará un script con instrucciones paso a paso para reproducir cada fase del análisis.

En el ámbito del trading, las regresiones pueden revelar patrones ocultos en los datos históricos, lo que puede llevar a la generación de señales de compra y venta más informadas. Además, investigaremos cómo las regresiones logísticas pueden ser útiles en la clasificación de eventos financieros y la predicción de movimientos del mercado.

1.1 Motivación

Automatizar herramientas de análisis utilizadas en el mercado financiero, como el análisis fundamental y técnico, que se enfoca en el análisis de los precios en la bolsa de valores y sus volúmenes de venta para predecir su comportamiento e inferir probabilidades a favor del inversor.

1.2 Objetivo

A aquellos objetivos de inversión validados previamente por un proceso NLP se les aplicará un análisis de pronóstico con regresiones. Se aplicarán diversos modelos de regresión (se puede aplicar regresión logística) y otros pronósticos al histórico de la base de datos.

2 Fuente de datos y selección de datos

2.1 Fuente de datos

Se hizo un pre-procesamiento y limpieza de los datos en la asignatura de minería de datos en de la cual se puede destacar lo siguiente:

Se unieron 7 bases de datos (6 de yahoo finance, y 1 de twitter).

Inicialmente, entre todos los conjuntos de datos se tienen 10.574 filas. Entre las cuales, 2574 filas pertenecen a las bases de datos de las 6 empresas con 7 columnas cada empresa, y 8000 filas en la base de datos de twitter con 99 columnas cada fila.

La unificación de las bases de datos de las 6 empresas tiene como resultado 625 filas que incluye las columnas Open, High, Low y Close de cada empresa entre el año 2021 y el año 2023. Estas bases de datos se mezclaron usando cómo indice la fecha, y se generó como resultado una base de datos de 625 filas y 24 columnas.

Después de unir la base de datos de twitter y la de las empresas, al final del preprocesamiento y calidad de los datos se obtuvo una tabla .csv con 5838 registros y 48 columnas o atributos.

2.2 Selección de datos (variables)

Hay 5 variables independientes relacionadas al precio, y son el precio de cierre del día para cada una de las 5 empresas.

Hay 5 variables independientes relacionadas a la demanda y son el volumen de ventas del día para cada una de las 5 empresas.

A pesar de que la base de datos tiene 48 atributos, dichas variables independientes son suficientes para desarrollar las regresiones deseadas.

Las regresiones servirán para añadir un criterio de decisión de compra o venta al logaritmo de trading que se busca desarrollar. Esa regresión se validará con un proceso NLP (en estudios posteriores a este), para validar si los pronósticos de regresión coinciden con las recomendaciones de trading ubicadas en la base de datos a la que se aplicó minería de datos.

¿Qué datos son discretos y cuáles continuos?

Son 48 campos diferentes, de los cuales no se utilizarán todos, por lo cual se clasifican a continuación los más útiles que son tipo float (valores continuos), int (valores enteros), datetime (fecha).

Date (datetime)	Close_AAPL (float)	Volume_AAPL (int)	Close_COIN (float)	Volume_COIN (int)
5/6/2022	157,279999	116124600	103,739998	9026500

Close_GEHC (float)	Volume_GEHC (int)	Close_RUN (float)	Volume_RUN (int)	Close_TMUS (float)	Volume_TMUS (int)
76,900002	1751400	23,41	9343700	126,800003	4701900

Los anteriores atributos son necesarios como variables independientes que se usarán para calcular los valores de regresión.

Los demás atributos de la base de datos de minería que no se explican acá, se pueden ver en el anexo 1. Estos atributos no tienen mucho valor para las regresiones, pero si sirven para un proceso de validación de tendencias NLP posterior a las regresiones.

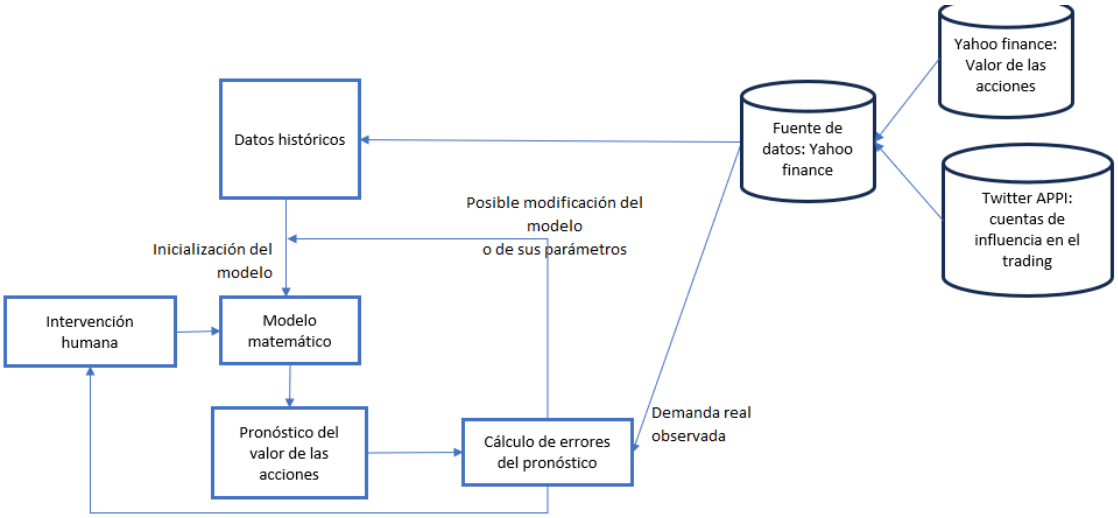
A continuación, con se describen los atributos útiles para la regresión que se pueden encontrar en la base de datos pre-procesada:

Date: Esta columna representa la fecha en la que se registraron los precios en yahoo finance. Permite alinear la fecha de inversión con la fecha de sugerencia de tendencia alcista en twitter para el posterior proceso de validación NLP.

Close: representa el precio de cierre de las acciones de la empresa al final del día de negociación. Hay uno precio de cierre para cada una de las empresas (5 empresas).

Volume: indica el volumen de acciones negociadas durante el día. Representa la cantidad total de acciones compradas y vendidas para cada una de las empresas (5 empresas).

Los modelos de pronósticos suelen hacer parte de un sistema que permite el análisis de los datos, en este caso, se ve de la siguiente forma:



De lo cual se puede observar que de las fuentes de datos se extraen los datos necesarios. De yahoo finance los valores de las acciones de distintas entidades u activos financieros y de twitter appi, se extraen los valores de los campos que permiten validar la credibilidad del tweet y analizar el texto con un análisis de sentimiento para determinar si existe intención de compra o inversión en la acción a analizar.

Es interesante agregar un análisis de demanda de las acciones basado en el volumen de los datos proporcionado por yahoo finance, ya que esto nos permitirá predecir a partir de la demanda de acciones cuando subirán y cuando bajarán los precios.

3 REGRESIÓN LINEAL Y REGRESIÓN POLINÓMICA

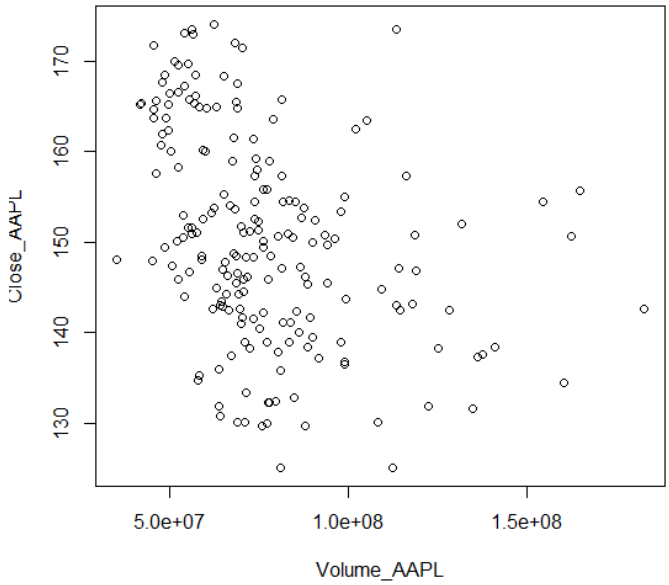
En la actividad realizada, se llevan a cabo las siguientes acciones:

Las columnas en la base de datos se ven así en R:

```
> print(non_numeric_volume)
 [1] X_id                  user.followers_count  user.friends_count
 [4] user.favourites_count retweet_count         favorite_count
 [7] created_at            text                  user.name
[10] user.screen_name      user.description      user.url
[13] new_date              Fecha_tradeo          Open_AAPL
[16] High_AAPL             Low_AAPL              Close_AAPL
[19] Adj.Close_AAPL        Volume_AAPL           Open_COIN
[22] High_COIN             Low_COIN              Close_COIN
[25] Adj.Close_COIN        Volume_COIN           Open_GEHC
[28] High_GEHC             Low_GEHC              Close_GEHC
[31] Adj.Close_GEHC        Volume_GEHC           Open_RUN
[34] High_RUN              Low_RUN               Close_RUN
[37] Adj.Close_RUN         Volume_RUN            Open_TMUS
[40] High_TMUS             Low_TMUS              Close_TMUS
[43] Adj.Close_TMUS        Volume_TMUS           year
[46] month                 weekday               day
<0 rows> (or 0-length row.names)
```

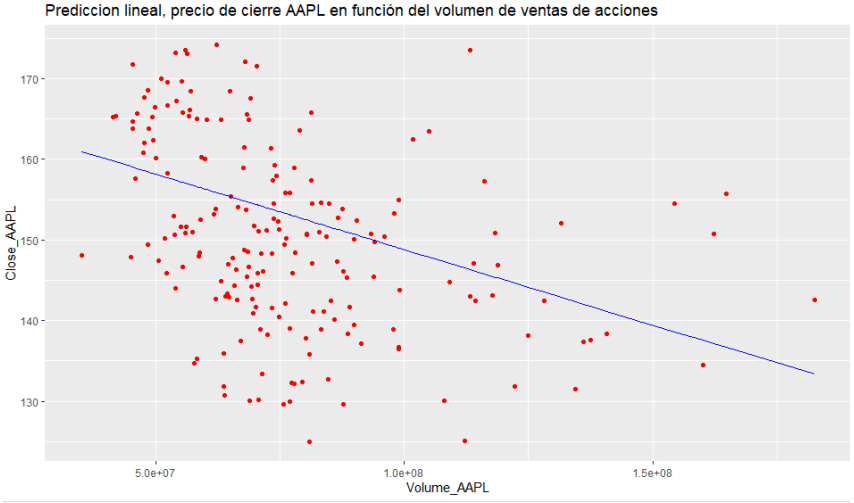
Todas sirven para complementar el proceso de automatización y NLP de las inversiones con un análisis de regresiones que posteriormente se usarán en algoritmos de trading en diferentes plataformar. Por ello se dejan los parámetros útiles en futuros trabajos junto con los útiles en esta asignatura. Siempre se usarán los valores Close, Volume y new_date.

- Se establece una URL como fuente de datos desde un archivo CSV alojado en GitHub, se cargan la bibliotecas necesarias, se importa el conjunto de datos desde el archivo .csv del repositorio git hub.
- Se verifica la naturaleza numérica de las columnas "Close_AAPL" y "Volume_AAPL" y se comprueba si existen valores faltantes (NA).
- Se crea un gráfico de dispersión utilizando los datos de "Volume_AAPL" y "Close_AAPL" con la función "ggplot", lo que resulta en un gráfico de dispersión original.

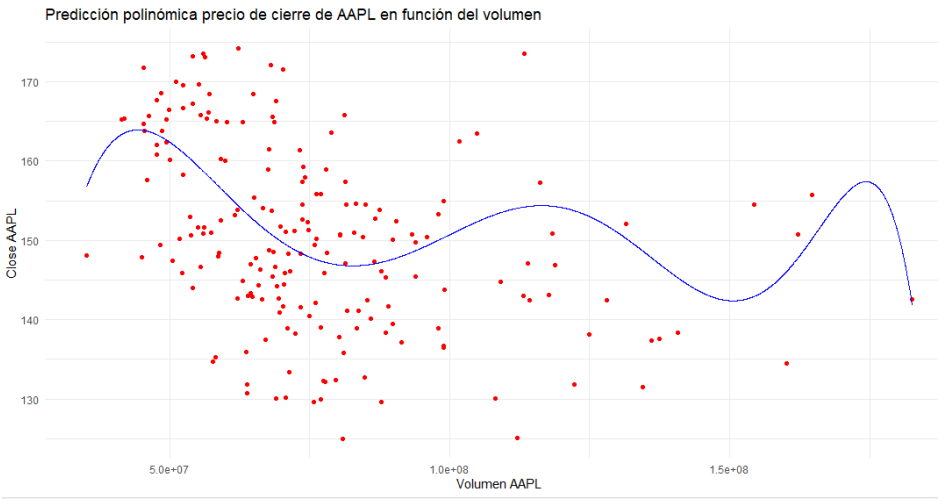


Aquí podemos ver que hay cierta tendencia, pero no es muy clara, y por ello hay que hacer análisis de regresión. Aparentemente, tiene tendencia descendiente

- Se identifican y muestran las filas con valores no numéricos en las columnas "Close_AAPL" y "Volume_AAPL" y se hace una validación del tipo de datos en las columnas actualmente.
- Se ajusta un modelo de regresión lineal utilizando la columna "Volume_AAPL" como variable independiente y "Close_AAPL" como variable dependiente. Luego, se realizan predicciones utilizando este modelo y se visualiza el modelo lineal junto con los datos originales en un gráfico.



  - Se puede observar una tendencia inversamente proporcional entre el aumento de Volume_AAPL y Close_AAPL. **Esto indica que hay que agregar más variables al estudio, puesto que el comportamiento normal del precio es aumentar cuando aumenta la demanda. Es por esto que después se programará un proceso NLP con análisis de tendencias en redes sociales de trading o noticias financieras.**
- Se agrega información polinómica a la base de datos creando columnas con potencias de la variable "Volume_AAPL" (hasta la cuarta potencia), se ajusta un modelo de regresión polinómica utilizando las variables originales de "Volume_AAPL" y sus potencias como variables independientes. Se utiliza la función "poly()" para transformar "Volume_AAPL" en una variable polinómica de grado 8 y se ajusta el modelo.
- Se crea un rango de valores para la variable "Volume_AAPL" para realizar predicciones con el modelo polinómico y se generan predicciones para estos valores. Finalmente, se crea un gráfico que muestra el modelo polinómico junto con los datos originales, utilizando las predicciones del modelo para trazar la línea polinómica.



- Luego, se muestra un resumen del modelo

```
> summary(regresion_polinomica)

Call:
lm(formula = Close_AAPL ~ poly(Volume_AAPL, degree = 8), data = base_datos)

Residuals:
    Min      1Q  Median      3Q     Max
-28.980  -6.991   1.548   7.891  21.862

Coefficients:
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                      154.558      0.136 1136.201  < 2e-16 ***
poly(Volume_AAPL, degree = 8)1 -329.524     10.392  -31.710  < 2e-16 ***
poly(Volume_AAPL, degree = 8)2  242.959     10.392   23.380  < 2e-16 ***
poly(Volume_AAPL, degree = 8)3 -172.811     10.392  -16.629  < 2e-16 ***
poly(Volume_AAPL, degree = 8)4  -29.773     10.392   -2.865 0.004184 **
poly(Volume_AAPL, degree = 8)5  135.187     10.392   13.009  < 2e-16 ***
poly(Volume_AAPL, degree = 8)6  -70.035     10.392   -6.739 1.74e-11 ***
poly(Volume_AAPL, degree = 8)7  -36.818     10.392   -3.543 0.000399 ***
poly(Volume_AAPL, degree = 8)8  -14.592     10.392   -1.404 0.160306
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.39 on 5827 degrees of freedom
Multiple R-squared:  0.2618,    Adjusted R-squared:  0.2607
F-statistic: 258.3 on 8 and 5827 DF,  p-value: < 2.2e-16
```

Observando el gráfico y el resumen estadístico, se puede concluir que:

Coeficientes de los Términos Polinómicos:

Los parámetros relevantes para esta conclusión son los coeficientes de los términos polinómicos (parámetro "Coefficients").

Los términos polinómicos de segundo, tercer, quinto y sexto grado tienen coeficientes significativos (indicados con asteriscos en la columna "Pr(>|t|)"). Esto sugiere que estas potencias de Volume_AAPL contribuyen de manera significativa a la predicción de Close_AAPL.

Los términos polinómicos de primer, cuarto, séptimo y octavo grado no son significativos, ya que sus valores p son mayores que 0.05. Esto indica que estas potencias de Volume_AAPL no contribuyen de manera significativa a la predicción.

R-**squared:**

El parámetro relevante para esta conclusión es el coeficiente de determinación (R-cuadrado).

El valor de R-cuadrado es 0.2618, lo que significa que aproximadamente el 26.18% de la variabilidad en Close_AAPL se explica mediante el modelo. Este valor relativamente bajo indica que el modelo no captura una gran parte de la variabilidad en la variable dependiente.

F-statistic:

El parámetro relevante es la estadística F y el valor p asociado.

El valor F es 258.3, y el valor p es prácticamente cero ($p < 2.2e-16$). Esto indica que el modelo en su conjunto es significativo y que al menos uno de los coeficientes es distinto de cero.

En resumen, el modelo de regresión polinómica utilizado parece explicar solo una fracción modesta de la variabilidad en Close_AAPL, ya que el R-cuadrado es relativamente bajo. Además, no todos los términos polinómicos son significativos en la predicción, lo que sugiere que la relación entre Close_AAPL y Volume_AAPL es más compleja de lo que se refleja en este modelo.

4 REGRESIÓN MULTILINEAL

este script realiza un análisis financiero que incluye preprocesamiento de datos, cálculo de rendimientos, ajuste de un modelo de regresión múltiple y visualización de relaciones entre acciones y el rendimiento del portafolio. También evalúa la correlación entre los rendimientos de las acciones y el rendimiento del portafolio. Todo para un portafolio compuesto por tres acciones (AAPL, COIN y GEHC). Aquí está el resumen de las acciones clave:

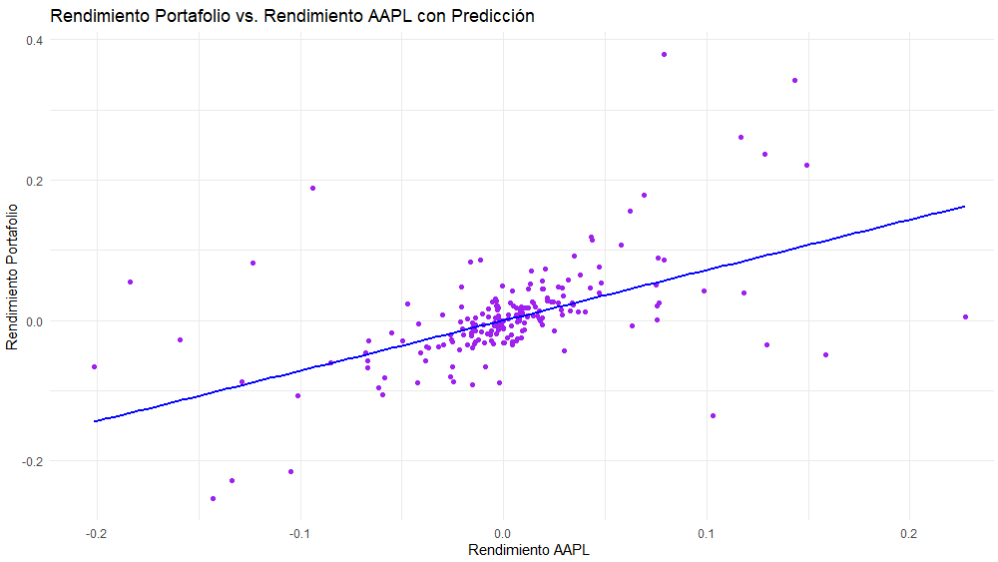
Se realiza el cálculo de rendimientos diarios, se definición de pesos del portafolio y se definen los pesos del portafolio para cada acción (por ejemplo, 40% en AAPL, 30% en COIN y 30% en GEHC).

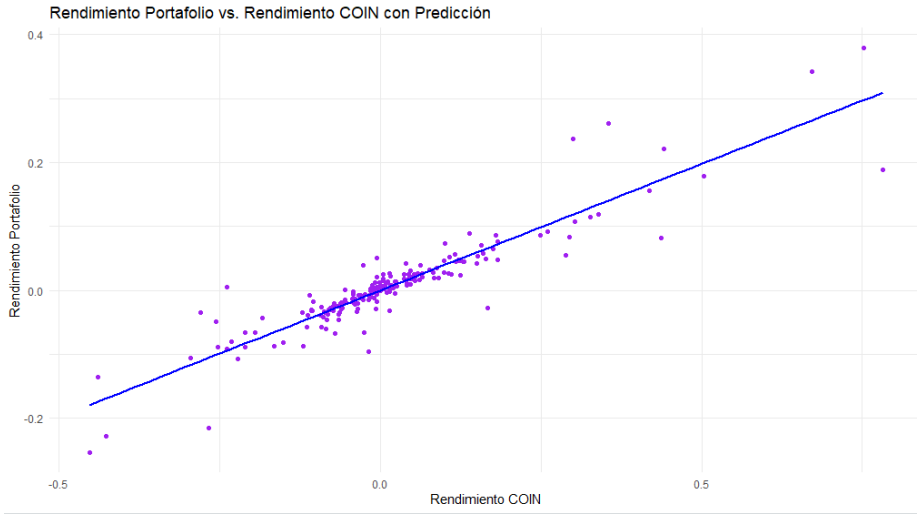
Ajuste de un Modelo de Regresión Múltiple:

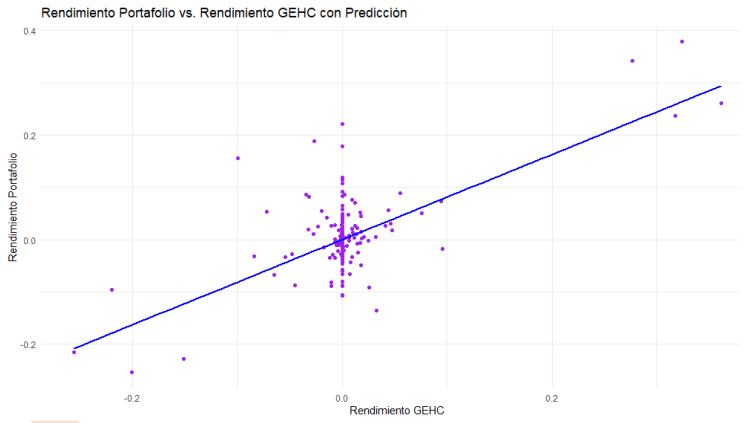
Se ajusta un modelo de regresión lineal múltiple para predecir el rendimiento diario del portafolio en función de los rendimientos diarios de las acciones individuales.

Visualización de Rendimientos y Predicciones:

Se crean gráficos de dispersión y líneas de regresión para visualizar la relación entre los rendimientos diarios del portafolio y cada acción individual (AAPL, COIN y GEHC).

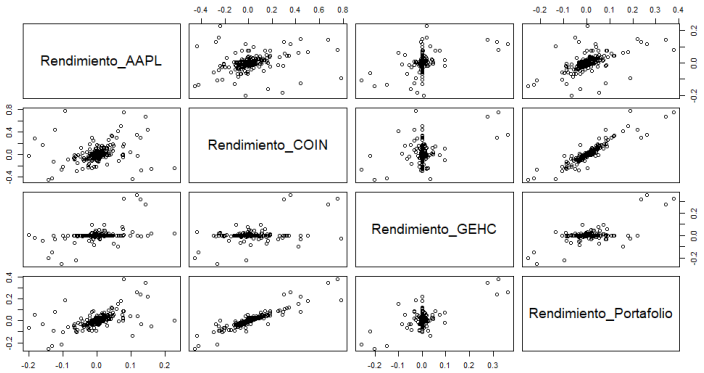




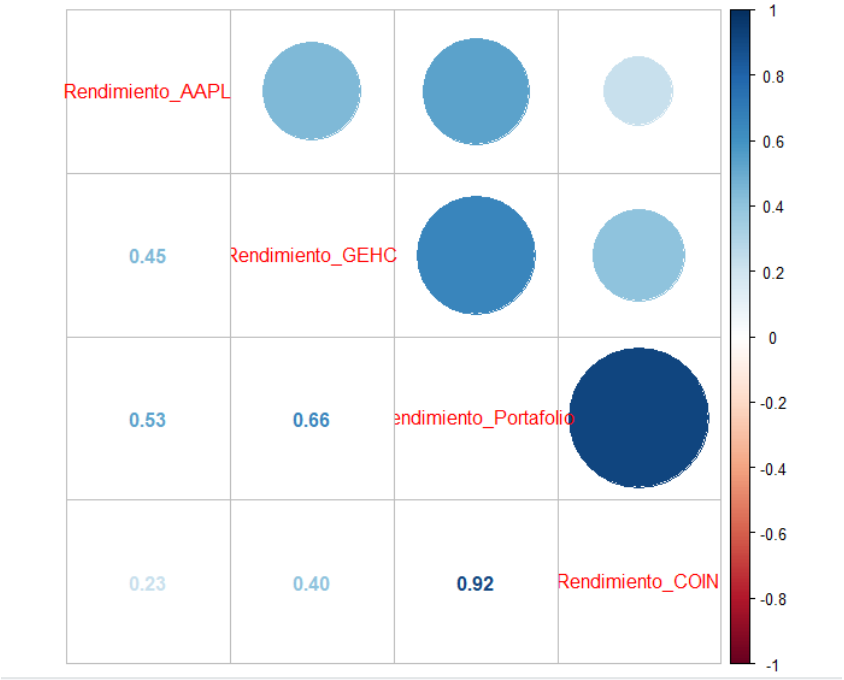


En los gráficos anteriores se puede observar la relación entre el rendimiento total del portafolio y cada uno de los precios de cierre de las 3 empresas

La correlación se puede ver de forma gráfica y en matriz de correlación así:



Matriz de Correlación:



Se calcula y muestra una matriz de correlación para evaluar las relaciones entre los rendimientos diarios de las acciones y el rendimiento del portafolio.

El resumen del modelo de regresión muestra lo siguiente:

```
> # Ver un resumen del modelo
> summary(modelo)

Call:
lm(formula = Rendimiento_Portafolio ~ Rendimiento_AAPL + Rendimiento_COIN +
    Rendimiento_GEHC, data = base_datos)

Residuals:
      Min        1Q    Median        3Q       Max
-0.034760  0.000024  0.000024  0.000024  0.027183

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2.362e-05  1.661e-05  -1.422    0.155
Rendimiento_AAPL  3.366e-01  1.895e-03 177.590   <2e-16 ***
Rendimiento_COIN  3.294e-01  5.912e-04 557.227   <2e-16 ***
Rendimiento_GEHC  2.989e-01  1.843e-03 162.194   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001269 on 5831 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9909,    Adjusted R-squared:  0.9908
F-statistic: 2.106e+05 on 3 and 5831 DF,  p-value: < 2.2e-16
```

Del resumen del modelo de regresión lineal, se pueden extraer varias conclusiones clave:

Coeficientes Significativos: Todos los coeficientes de las variables predictoras (Rendimiento_AAPL, Rendimiento_COIN y Rendimiento_GEHC) son altamente significativos, como indicado por los códigos de significancia '***'. Esto sugiere que cada una de estas variables tiene un impacto estadísticamente significativo en el rendimiento del portafolio.

R-cuadrado Alto: El valor del R-cuadrado múltiple es muy alto (0.9909), lo que significa que el modelo es altamente efectivo para explicar la variabilidad en el rendimiento del portafolio. Aproximadamente, el 99.09% de la variabilidad en el rendimiento del portafolio se puede explicar por las variables predictoras incluidas en el modelo.

Modelo Globalmente Significativo: El estadístico F es extremadamente alto (2.106e+05), y el p-valor asociado es prácticamente cero (< 2.2e-16). Esto indica que el modelo en su conjunto es altamente significativo y que al menos una de las variables predictoras tiene un efecto significativo en el rendimiento del portafolio.

Error Estándar Bajo: El error estándar residual es muy bajo (0.001269), lo que sugiere que las predicciones del modelo son precisas en general.

En resumen, el modelo de regresión lineal es estadísticamente sólido y demuestra que las variables de rendimiento de acciones utilizadas son altamente significativas para predecir el rendimiento del portafolio. Además, el modelo tiene un alto poder explicativo, lo que lo hace útil para comprender y predecir el comportamiento del portafolio en función de las acciones individuales.

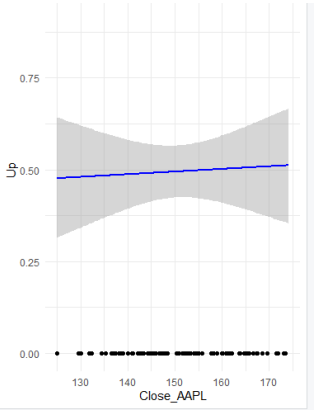
En resumen, el modelo de regresión lineal muestra que las variables de rendimiento de las acciones (AAPL, COIN y GEHC) son altamente significativas para predecir el rendimiento del portafolio, y el modelo explica una gran parte de la variabilidad en el rendimiento del portafolio (aproximadamente el 99.09%).

5 Regresión logística

```
base_datos3 <- base_datos2[, c("Up", "Close_AAPL")]

# Muestra los primeros 10 registros de los nuevos datos
head(base_datos3, 200)
```

```
> prediccion_logistica = predict(object = regresion_logistica, newdata = base_datos3, type = "response")
> prediccion_logistica
        9        17        24        26        33
0.4959837 0.4891412 0.4924181 0.4853902 0.4855630
       36        42        43        46        49
0.4895373 0.4953714 0.4911721 0.4917194 0.4935705
       57        67        77        89       105
0.4891988 0.4852319 0.4814538 0.4820869 0.4801446
      113       116       125       132       153
0.4812236 0.4843250 0.4860525 0.4884932 0.4854550
      156       158       161       169       172
0.4849367 0.4865276 0.4884212 0.4918707 0.4923676
      180       192       203       210       213
0.4908047 0.4915178 0.4933977 0.4946222 0.4923893
      219       226       236       243       244
0.4952201 0.4983539 0.4974461 0.4966249 0.4956523
      250       257       258       269       275
0.4997947 0.5035121 0.5027917 0.5017111 0.5058893
      277       287       293       297       303
```

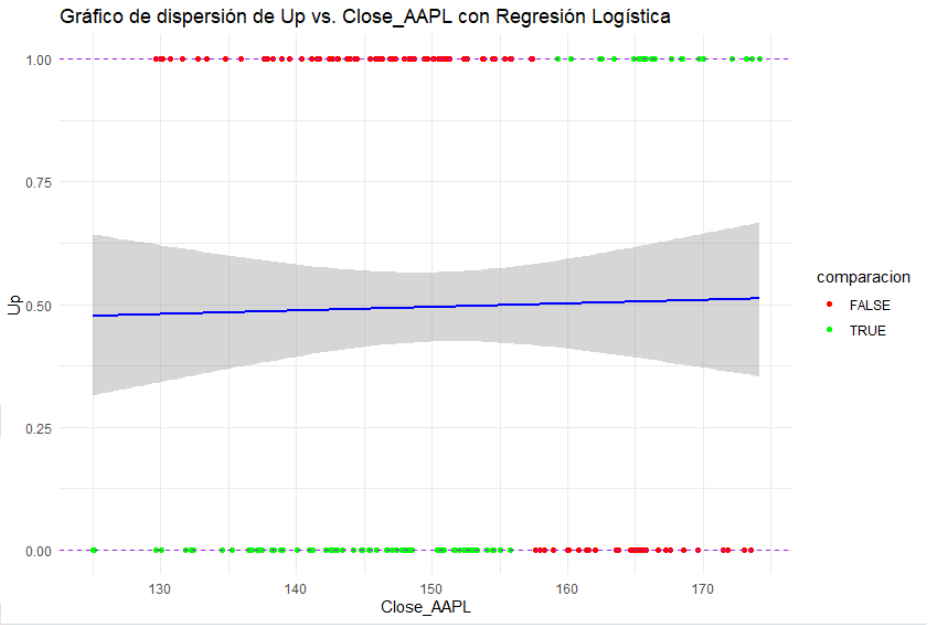


a razón por la que todos los valores de la predicción están cerca de 0.5 es debido a la forma en que se realiza la predicción en un modelo de regresión logística.

En un modelo de regresión logística, el resultado es la probabilidad de que la variable dependiente (en este caso, "Up") sea igual a 1 (o "éxito"). La probabilidad se encuentra en el rango de 0 a 1. Cuando obtienes valores cercanos a 0.5 en la predicción, significa que el modelo no está seguro de si la variable dependiente será 1 o 0, y está asignando una probabilidad cercana al 50% a ambas posibilidades.

Para obtener predicciones más seguras o distintas de 0.5, puedes ajustar el umbral de clasificación. Por ejemplo, puedes considerar que cualquier probabilidad por encima de 0.5 se clasifica como 1 y cualquier probabilidad por debajo de 0.5 se clasifica como 0. Esto dependerá de tus necesidades específicas y de cómo deseas interpretar las predicciones.

En resumen, la predicción cercana a 0.5 en un modelo de regresión logística indica incertidumbre o falta de confianza en la clasificación de las observaciones en una categoría específica. Puedes ajustar el umbral si deseas predicciones más claras.



En el código proporcionado, el usuario está llevando a cabo una comparación entre las predicciones y la columna "Up" en los datos. La columna "Up" parece indicar si el precio de cierre (Close_AAPL) ha subido (1) o no (0).

Los resultados muestran que cuando "Up" es igual a 0 y el valor de "Close_AAPL" es mayor que 160, las predicciones son 0, lo que significa que no se predice un aumento, y la comparación es verdadera (True). Esto indica que el modelo acierta al predecir cuando el precio de cierre no sube y supera los 160.

Por otro lado, cuando "Up" es igual a 0 y el valor de "Close_AAPL" es menor que 160, las predicciones son 1, lo que indica que se predice un aumento, y la comparación es falsa (False). Esto señala que el modelo no acierta en estos casos.

La razón por la que todos los puntos con "Up" igual a 0 y "Close_AAPL" mayor que 160 aparecen en rojo como Falsos en la gráfica se debe a que el modelo no está realizando predicciones precisas en estas situaciones según los datos disponibles.

Para mejorar la capacidad predictiva en estos casos, se sugiere revisar y ajustar el modelo de regresión logística, considerar posibles características adicionales o modificar los hiperparámetros del modelo. También sería beneficioso explorar si existe información adicional que pueda ayudar a mejorar las predicciones en estas circunstancias particulares

6 ANEXOS

6.1 Anexo 1: visualización de los campos existentes en la base de datos descargada desde el datalake de twitter.

Los siguientes campos con ejemplos de sus correspondientes valores, provienen de la base de datos generada con procesamiento de datos en la clase de minería de datos:

_id	user.followers_count	user.friends_count	user.favourites_count	retweet_count	favorite_count	created_at
645a831bf55dd5c337018609	234157	771	4211	10	54	Fri May 06 10:09:07 +0000 2022

text	user.name	user.screen_name	user.description	user.url
This is an important part of the Archegos / GSX / $GOTU story. As GAX squeezed higher, desks told us there was a la… https://t.co/bPBYLz2cN0	MuddyWatersResearch	muddywatersre	Activist short seller, skeptic, First Amendment advocate, foot soldier in the Global War to Defend Truth	https://t.co/bZSmfKyo8x

Date (datetime)	Open_AAPL (float)	High_AAPL (float)	Low_AAPL ((float)	Close_AAPL (float)	Adj Close_AAPL (float)	Volume_AAPL (int)	Open_COIN (float)	High_COIN (float)	Low_COIN (float)
5/6/2022	156,009995	159,440002	154,179993	157,279999	156,34642	116124600	112,5	112,5	100,25

Close_COIN (float)	Adj Close_COIN (float)	Volume_COIN (int)	Open_GEHC (float)	High_GEHC (float)	Low_GEHC (float)	Close_GEHC (float)	Adj Close_GEHC (float)	Volume_GEHC (int)
103,739998	103,739998	9026500	76,860001	77,540001	76,264999	76,900002	76,870926	1751400

High_RUN (float)	Low_RUN (float)	Close_RUN (float)	Adj Close_RUN (float)	Volume_RUN (int)	Open_TMUS (float)	High_TMUS (float)	Low_TMUS (float)	Close_TMUS (float)
25,200001	22,66	23,41	23,41	9343700	128,259995	129,490005	125,089996	126,800003