IA EN LOS MERCADOS FINANCIEROS. CONSECUENCIAS Y CASO PRACTICO

by

Arnau Muns Orenga

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DEGREE

in

Field

Approved:

Chair Person Dr. Salvador Torra	Committee Member 1 Committee Member

UNIVERSITAT DE BARCELONA
Barcelona, España

2018

DEDICACIÓN

Este trabajo está dedicado a mi familia.

RESUMEN

Resumen

Palabras clave

Stock, Inteligencia artificial, Aprendizaje automático, Modelo predictivo, precios OHLC

ÍNDICE

LISTA DE TABLAS

LISTA DE FIGURAS

CAPÍTULO 1
Introducción

La inteligencia artificial, comúnmente llamada AI por sus siglas en inglés…

Intentar predecir el precio de un stock ha sido una materia objeto de estudio desde los últimos años. Debido a la complejidad de las series temporales de los precios una nueva rama de investigación se está desarrollando rápidamente. Ésta rama de investigación gira en torno a la idea de intentar predecir la *dirección* del stock en cuestión en vez de intentar predecir el precio exacto. Estos modelos de predicción de la dirección del movimiento de los stocks, llamados SMD por sus siglas en inglés (Stock Movement Direction), consiguen alcanzar un rendimiento predictivo suficiente como para utilizarlos como complemento al análisi fundamental a la hora de tomar una decisión de inversión.

siempre tenemos que poner un justifying despues de un center o nos va a centrar todo el documento

CAPÍTULO 2

Metodología

La metodología que se va a seguir en el presente trabajo es la siguiente:

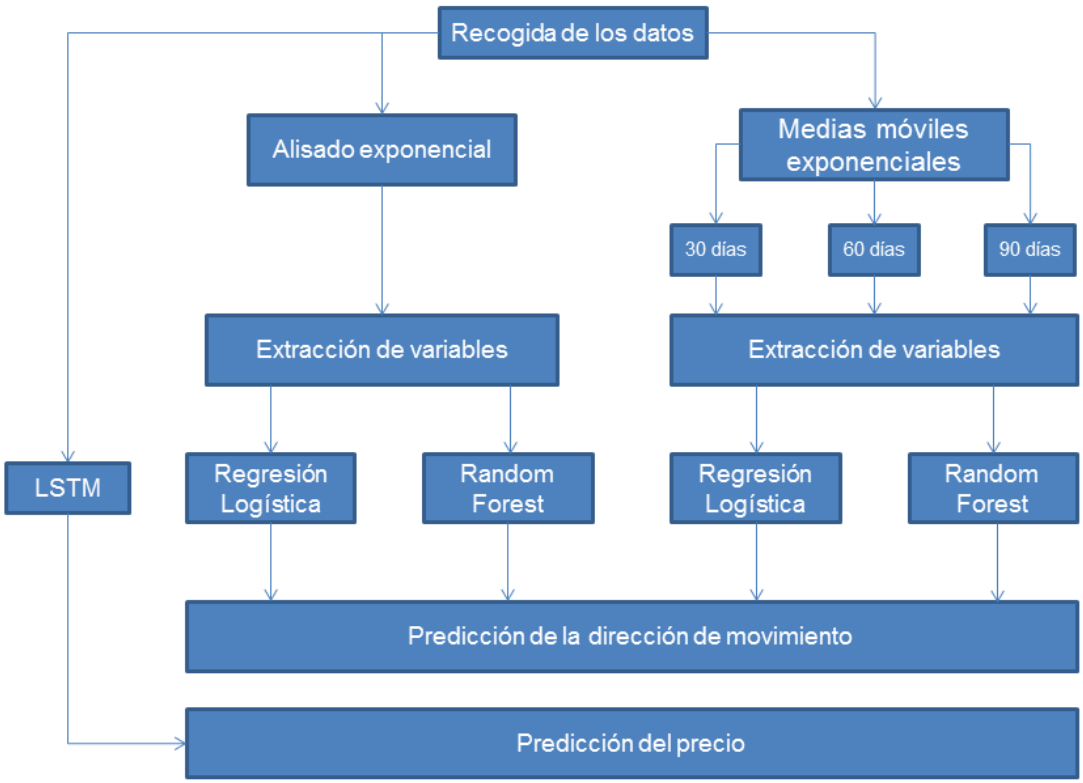


Figura 2.1: Metodología propuesta

CAPÍTULO 3

Consecuencias de la implantación de la IA en los mercados financieros

Tells us more.

3.1 Contextualización

no se que pasa pero el titulo no sale bien

3.2 Evolución histórica

3.3 Consecuencias económicas

CAPÍTULO 4

Caso práctico: modelización predictiva de la dirección de un stock

Explicar la motivación para llevar a cabo un caso práctico. Explicar el tipo de modelos que se van a desarrollar y cómo encajan éstos en la situación actual de aplicación de la IA en los mercados financieros.

4.1 Investigaciones previas

CAPÍTULO 5
Base de datos

5.1 Obtención y descripción

- Cómo se obtienen. Procedimiento y fuente de los datos. Explicar las distintas tablas que conforman la base de datos.
- Explicar cuantas empresas forman el portfolio. Explicar con qué criterio se seleccionan las empresas que forman el portfolio (Elección propia en función de los portfolios de distintos gurús del mundo de las financias. i.e. Warren Buffet)
- Descripción de los datos (variables + estadística descriptiva)

Obtención

Los datos con los que se va a realizar el presente trabajo se obtienen a través del paquete `quantmod`. Este paquete se diseñó para assistir a los *traders* quantitativos en el desarrollo, evaluación y puesta en funcionamiento de modelos de *trading* basados en la estadística. (revisar, cita (pckg info)). Concretamente, se utiliza la función `getSymbols` para obtener los datos ya que permite cargar y manejar `Symbols` en un ambiente especificado (cita (funs help)). Las posibles fuentes de datos son: Yahoo; Google, aunque actualmente no funciona; Alphavantage, con la ventaja de que podemos obtener datos intra-dia; MySQL, FRED, etc. (cita). Para obtener los datos se procede a utilizar *Yahoo Finance* como la fuente.

La metodología para la búsqueda de los distintos stocks con los cuales entrenar los modelos predictivos empieza con la idea de querer encontrar empresas representativas dentro de distintos sectores estratégicos y con amplio impacto en la economía real. Además se pretende que los stocks utilizados tengan rentabilidad financiera con el objetivo de llevar el presente trabajo lo más cerca posible de una situación real de inversión. Para ello se utiliza el portfolio a fecha 01/01/2019 del gran *gurú* de las finanzas Warren Buffet ("Warren buffett," n.d.). Dentro de las numerosas empresas presentes en este portfolio se escogen los 3 *stocks* del NYSE (bolsa de nueva york) y una del NASDAQ.

Simbolo	Nombre	Bolsa
AAPL	Apple Inc.	NASDAQ
WFC	Wells Fargo & CO	NYSE
KO	Coca-Cola Company	NYSE
AXP	American Express CO	NYSE

Tabla 5.1: Stocks utilizados

Al hecho de que los *stocks* que forman la base de datos formen parte del portfolio de uno de los *gurús* del mundo de las finanzas y la inversión, se le suma también el hecho de que las empresas escogidas son empresas multinacionales con un gran capital que ocupan

una posición destacada sus respectivos mercados, siendo representativas de cada uno de los sectores en los cuales operan.

Coca - cola (explicar un poco la compañía) etc etc

Descripción

Después de utilizar la función `getSymbols` se obtiene una tabla para cada stock con un formato estandarizado. Se obtienen 5 series temporales con periodicidad diaria que hacen referencia a los precios de **apertura**, **cierre**, **máximo**, **mínimo** y **volúmen**. Además se incluye el precio ajustado pero esta variable no será utilizada en este trabajo. Todas las variables presentes en la base de datos utilizada son numéricas.

Los datos utilizados en este trabajo compreden el período 2000 - ?, siendo el día `2000-01-01` la primera observación de cada serie temporal. La última observación de la base de datos es (revisar)

La función `str` permite visualizar fácilmente la estructura y formato que presentan en R las distintas tablas de la base de datos inicial. En el siguiente output se muestra como ejemplo la empresa Coca-Cola Company.

```
## 'data.frame':    4778 obs. of  6 variables:
##  $ KO.Open    : num  29 28.2 28.2 28.5 28.9 ...
##  $ KO.High    : num  29 28.4 28.7 28.8 30.4 ...
##  $ KO.Low     : num  27.6 27.8 28 28.3 28.9 ...
##  $ KO.Close   : num  28.2 28.2 28.5 28.5 30.4 ...
##  $ KO.Volume  : num  10997000 7308000 9457400 7129200 11474000 ...
##  $ KO.Adjusted: num  12.5 12.5 12.6 12.6 13.4 ...
```

Tablas descriptivas

Se explora descriptivamente los datos analizando los estadísticos descriptivos . Para cada empresa, se obtienen los distintos estadísticos de cada una de las variables descritas previamente. Para ello se elaboran 4 tablas que hacen referencia a los distintos estadísticos, calculados sobre los distintos precios y el volúmen, para una misma empresa.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Open	18.55	23.49	28.75	31.43	40.4	50.82
High	18.8	23.72	29.01	31.66	40.6	50.84
Low	18.5	23.26	28.47	31.19	40.15	50.25
Close	18.54	23.5	28.77	31.44	40.39	50.51
Volume	2.147M	9.821M	12.972M	14.692M	17.49M	124.169M

Tabla 5.2: Estadísticos descriptivos para los distintos precios de Coca-Cola Company

Como se puede apreciar en la tabla 5.2 el rango que han tomado los precios de cierre para la empresa Coca-Cola CO. en el período estudiado se mueve entre 18.54 y 50.51. El precio medio de cierre para el período estudiado es de 31.44 dólares.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Open	0.93	4.37	26.13	51.65	90.52	230.78
High	0.94	4.47	26.53	52.13	91.37	233.47
Low	0.91	4.21	25.66	51.12	89.72	229.78
Close	0.94	4.36	26.1	51.64	90.52	232.07
Volume	9.835M	52.133M	93.003M	119.542M	156.041M	1855.41M

Tabla 5.3: Estadísticos descriptivos para los distintos precios de Apple Inc.

Como se puede apreciar en la tabla 5.3 el rango que han tomado los precios de cierre para la empresa Apple Inc. en el período estudiado se mueve entre 0.94 y 232.07. El precio medio de cierre para el período estudiado es de 51.64 dólares, superior al de la empresa Coca-Cola CO.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Open	9.99	41.14	50.19	55.59	69.94	113.99
High	10.66	41.58	50.89	56.16	70.45	114.55
Low	9.71	40.5	49.72	55	69.56	112
Close	10.26	41.04	50.22	55.59	70.08	112.89
Volume	0.837M	3.85M	5.314M	6.979M	7.965M	90.337M

Tabla 5.4: Estadísticos descriptivos para los distintos precios de American Express CO.

Como se puede apreciar en la tabla 5.4 el rango que han tomado los precios de cierre para la American Express CO. en el período estudiado se mueve entre 10.26 y 112.89. El precio medio de cierre para el período estudiado es de 55.59 dólares, ligeramente superior al de la empresa Apple Inc..

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Open	8.65	25.95	31.19	35.19	45.82	65.89
High	8.94	26.29	31.48	35.56	46.23	66.31
Low	7.8	25.54	30.84	34.82	45.47	65.66
Close	8.12	25.91	31.2	35.2	45.97	65.93
Volume	1.774M	9.129M	15.235M	23.412M	27.088M	478.737M

Tabla 5.5: Estadísticos descriptivos para los distintos precios de Wells Fargo and CO.

Como se puede apreciar en la tabla 5.4 el rango que han tomado los precios de cierre para la empresa Wells Fargo and CO. en el período estudiado se mueve entre 8.12 y 65.93. El precio medio de cierre para el período estudiado es de 35.2 dólares, ligeramente superior al de Coca-Cola CO. inferior al de `AAPL` y `AXP`.

Visualización gráfica

Seguidamente se explora gráficamente el precio de cierre para cada una de las empresas, al ser éste la variable a partir de la cual se calculará la variable respuesta sobre la que hacer predicción. El objetivo de este tipo de análisis es poder visualizar el tipo de evolución que han presentado los distintos *stocks* durante el período de estudio.

En el siguiente gráfico se representa el precio de cierre de Coca-Cola Company para el período comprendido entre el 03/01/2000 y 28/12/2018.

Figura 5.1: Precio de cierre de Coca-Cola Company 03/01/2000 - 28/12/2018

En el siguiente gráfico se representa el precio de cierre de Apple Inc. para el período comprendido entre el 03/01/2000 y 28/12/2018.

Figura 5.2: Precio de cierre de Apple Inc. 03/01/2000 - 28/12/2018

En el siguiente gráfico se representa el precio de cierre de Wells Fargo and CO. para el período comprendido entre el 03/01/2000 y 28/12/2018.

Figura 5.3: Precio de cierre de American Express CO. 03/01/2000 - 28/12/2018

En el siguiente gráfico se representa el precio de cierre de Wells Fargo and CO. para el

período comprendido entre el 03/01/2000 y 28/12/2018.

Figura 5.4: Precio de cierre de Wells Fargo and CO. 03/01/2000 - 28/12/2018

A partir de esta exploración gráfica se descubre que 3 de las 4 empresas (KO, AXP y WFC) presentan un comportamiento relativamente similar. Como se puede observar en la figura 5.1, 5.3 y 5.4 la evolución de `KO, AXP y WFC` en el período estudiado presenta una tendencia, en general, creciente. Si se analiza en detalle el lector puede detectar 3 etapas claramente destacada en la evolución de estos stocks, siendo la primera de ellas distinta en el caso de Wells Fargo & CO. Estos 3 períodos claramente diferenciados son los correspondientes a los años 2000-2007, 2008-2010 y 2011-2018.

2000-2006

En el caso de las empresas Coca-Cola CO. y American Express CO. la primera etapa (2000 - inicios de 2006) corresponde a la relajación de los mercados financieros posterior al *boom de las .com*. La segunda mitad de los 90 fueron una época de expansión económica en la cual se produjo un boom financiero en USA (ver Urban Jermann, 2003) que se relajo a partir de finales de milenio.

2006-2008

Esta relajación fué seguida por la segunda etapa, un período de crecimiento económico y de los mercados financieros que se produjo a partir del 2006 y hasta el 2008. Esta etapa de

crecimiento fué previa a la gran recesión mundial que tuvo lugar a partir del año 2008 (ver Andrew K. Rose, 2011, DeLong (2009), P.R. Lane (2011) pp. 77-110 para más información)

Descriptiva financiera

CONTAR LAS VECES QUE EL PRECIO EN EL DÍA T ES MAYOR/INFERIOR A T-1 QUE EL PRECIO A BAJADO EN EL PRECIO SIN ALISAR. DESPUES CON EL PRECIO ALISADO Y COMPARANDO LA % DE DÍA

SABEMOS QUE ESTA VARIABLE SE PUEDE PREDECIR MUY BIEN A.K.A PODEMOS CLASIFICAR MUY BIEN CON INDICADORES TECNICOS LOS DÍAS QUE INCREMENTA O DECREMENTA EL PRECIO. ESTO PRUEBA QUE LOS INDICADORES TECNICOS SON UTILES A LA HORA DE HACER INVERSIONES PORQUE, ANALIZADOS CORRECTAMENTE, "PERMITEN" SABER SI EN UN DIA T EL PRECIO HABRÁ SUBIDO O NO RESPECTO A T-1. ESTA NO ES LA VARIABLE RESPUESTA QUE SE UTILIZA EN ESTE TRABAJO, AL UTILIZAR INDICADORES TECNICOS CALCULADOS CON EL PROPIO EN PRECIO EN T PARA PODER SABER SI EL PRECIO EN T HA SUBIDO RESPECTO T-1

Seguidamente se muestran las distintas tablas comparativas de los distintos precios obtenidos inicialmente. Éstas incluyen la media, la desviación típica o volatilidad del precio y el coeficiente de variación, el cual es un estadístico utilizado de manera frecuente para medir la volatilidad de un stock al ser una variable que no tiene en cuenta las unidades de medida. Esto significa que permite comparar distintos stocks en cuanto a volatilidad se refiere sin tener en cuenta la magnitud que éstos presenten.

Este tipo de visualización permite analizar más detenidamente cada uno de los precios y poder compararlos fácilmente entre todas las empresas. Es además una buena manera de analizar a priori la predictibilidad de los distintos precios al poder comparar los distintos coeficientes de variación. La idea de evaluar la predictibilidad de una serie temporal con el coeficiente de variación sugiere que es más fácil de predecir una serie temporal con un comportamiento más estable, esto es, con menos unidades de desviación típica por unidad de media (Gilliland, 2009).

Nombre	Media	Desv_std	CoV
Apple Inc.	51.64595	55.925416	1.0828616
Wells Fargo & CO	35.19440	11.806917	0.3354771
Coca-Cola Company	31.42690	8.652774	0.2753302
American Express CO	55.58608	21.112421	0.3798149

Tabla 5.6: Estadísticos descriptivos para el precio de apertura (Open).

COMENTAR TABLAS??

Nombre	Media	Desv_std	CoV
Apple Inc.	52.13167	56.386960	1.0816258
Wells Fargo & CO	35.56184	11.819827	0.3323739
Coca-Cola Company	31.66443	8.670834	0.2738351
American Express CO	56.16119	21.131090	0.3762579

Tabla 5.7: Estadísticos descriptivos para el precio máximo (High).

Nombre	Media	Desv_std	CoV
Apple Inc.	51.12417	55.450009	1.0846143
Wells Fargo & CO	34.82003	11.805231	0.3390357
Coca-Cola Company	31.18907	8.644319	0.2771585
American Express CO	55.00056	21.083165	0.3833264

Tabla 5.8: Estadísticos descriptivos para el precio mínimo (Low).

Nombre	Media	Desv_std	CoV
Apple Inc.	51.63773	55.925371	1.0830331
Wells Fargo & CO	35.19502	11.807079	0.3354758
Coca-Cola Company	31.43855	8.656529	0.2753476
American Express CO	55.58724	21.099819	0.3795803

Tabla 5.9: Estadísticos descriptivos para el precio de cierre (Close).

5.2 Procesamiento de los datos

Siguiendo la metodología propuesta en la figura 2.1, en primer lugar se procede a elaborar una transformación de todas las variables/precios iniciales. Esta transformación suaviza los datos con el objetivo de remover la variación aleatória y/o el ruido, haciendo que los modelos predictivos de dirección de movimiento elaborados posteriormente sean capaces de detectar más fácilmente una tendencia a largo plazo de los precios dentro del comportamiento de los mismos (Luckyson Khaidem, 2016, p. 4). El objetivo es, por un lado, crear la variable respuesta utilizando el precio de cierre alisado y, por otro, calcular los indicadores técnicos detallados en la sección 5.3 para, posteriormente, utilizarlos como variables predictoras.

Por un lado se utilizan las medias móviles exponenciales a 30, 60 y 90 días como nuevas representaciones de los precios. De esta manera se obtiene una nueva representación de los precios en la cual se ha removido la variación aleatória y el ruido. Seguidamente se muestran las gráficas para las medias móviles exponenciales del precio de cierre para la empresa Coca-Cola CO. con el objetivo de visualizar el alisado que se obtiene al aplicar este cálculo. A medida que se añaden más datos en el cálculo de las medias se obtiene un mayor alisado.

Por otro lado se aplica el mismo tipo de alisado exponencial aplicado por (Luckyson Khaidem, 2016, p. 4). El objetivo es el mismo que en el caso del cálculo de las medias móviles. El estadístico alisado exponencialmente de un serie temporal Y se puede calcular recursivamente, desde el momento en el que se disponga de dos observaciones, de la manera siguiente:

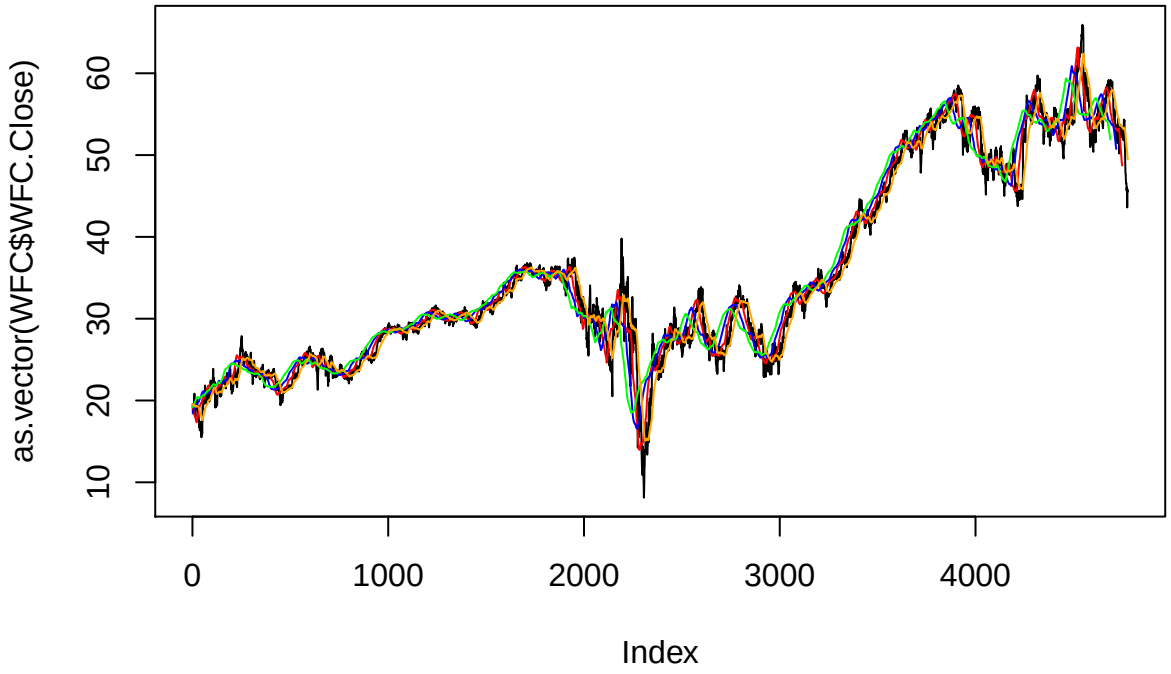
$$S_0 = Y_o$$

$$S_t = \alpha * Y_y + (1 - \alpha) * S_{t-1} \quad \forall \ t > 0 \qquad (1)$$

donde $0 < \alpha < 1$ es el factor de alisado. El método del alisado exponencial otorga más peso a las observaciones más recientes en tanto que hace decrecer exponencialmente el peso de las observaciones más antiguas. (cita articulo, formula internet). Valores elevados de α provocan que el nivel de alisado quede reducido, de manera que en el caso extremo en el que $\alpha = 1$ el estadístico exponencialmente alisado es igual que la observación. Es decir $S_t = Y_t \quad si\ \alpha = 1$. En el presente trabajo se aplica la fórmula anterior sobre cada uno de los precios y el volúmen con un factor de alisado $\alpha = 0.05$ para conseguir un alisado más pronunciado que en el caso de las medias móviles exponenciales.

comparativa formula vs 30, 60, 90, quien es mas liso, sentido economico etc. predecir la media de a 1 mes vista de EMA 30 es predecir si en 1 mes la media de los precios de ese mes comparada con las del mes que queda detrás ha subido o no

revisar, poner toda la grafica junta comparativa de todas las lineas a la vez



A su vez se presentan la siguiente tabla para poder visualizar la disminución de variabilidad que se obtiene tras aplicar las distintas medias móviles exponenciales al precio de cierre de las distintas empresas. La volatilidad se evalúa en este caso con el coeficiente de variación para permitir la comparabilidad:

Nombre	CoV_EMA30	CoV_EMA60	CoV_EMA90	CoV.exp.smooth.alpha.0.05
Apple Inc.	1.0799103	1.0746436	1.0686425	1.0852714
Wells Fargo & CO	0.3322648	0.3290268	0.3263450	0.3336280
Coca-Cola Company	0.2735308	0.2716672	0.2700638	0.2722442
American Express CO	0.3763897	0.3725283	0.3686093	0.3746762

Tabla 5.10: Comparativa coeficiente de variación entre las distintas medias móviles exponenciales

Una vez alisados los precios de los distintos *stocks* se procede a calcular tanto la variable respuesta como las variables predictoras, en este caso los indicadores técnicos, a partir de éstos.

5.3 Creación de variables

Variable respuesta

Después de aplicar ambos tipos de alisado sobre los datos se obtienen 4 bases de datos para cada una de las empresas, las cuales hacen referencia a los precios transformados con

las medias móviles exponenciales de 30, 60 y 90 días y utilizando la formula del alisado exponencial. En total la base de datos contiene ahora 16 tablas, 4 para cada empresa.

En este momento se define la variabe respuesta a partir de las distintas transformaciones del precio de cierre de siguiendo la fórmula siguiente:

$$Target_t = \begin{cases} Up & Close_{t+d} > Close_t \\ Down & Close_{t+d} < Close_t \end{cases} \tag{2}$$

donde:

- Close_t es el precio de cierre del activo en el día "t".

La variable respuesta en t toma el valor Up si el precio de cierre a día $t+d$ es superior al precio de cierre del día actual t. Del mismo modo, la variable respuesta en t toma el valor $Down$ si el precio de cierre en $t+d$ es inferior al precio de cierre del día actual t.

En la mayoría de la literatura revisada se construye la variable respuesta para la observación correspondiente al día t comparando el precio de cierre de se día con el precio de cierre de un día anterior $t-d$ (véase Yakup Kara, 2011, p. 5313, Masoud (2014) p. 610). Económicamente esta variable está midiendo si el precio actual ha tenido una evolución positiva, o no, respecto al valor que presentaba, por ejemplo, hace un mes. Posteriormente se calculan las variables predictoras, esto es, los indicadores técnicos, a partir de los precios OHLC medidos el día t. En este caso los modelos de aprendizaje automático obtienen, en general, buenos rendimientos y son capaces de clasificar una nueva observación. Sin embargo este modo de calcular la variable respuesta no brinda al inversor una herramienta útil para argumentar sus decisiones de compra o venta ya que para predecir la variable respuesta en t, que está calculada a partir del precio de cierre, se utilizan indicadores técnicos calculados, a su vez, con el precio de cierre (revisar poner indicadores solo utilizann precio de cierre). Es decir en el momento en el que se pueden obtener los directamente con el precio de cierre anterior para ver si el precio ha subido o no. Es por eso que en el presente trabajo se ha decidido construir la variable respuesta de una manera distinta, al ser el objetivo del mismo el de construir un modelo predictivo que sea útil a la hora de tomar una decisión de inversión. Por eso se ha decidido construir la variable respuesta como la construida en (Luckyson Khaidem, 2016, p. 4, Han (2012) p. 16 y Kim (2003) p. 4). Creando la variable a predecir de este modo, el inversor puede utilizar los precios en t para predecir la dirección que tomará el precio de ese *stock* al cabo de 1, 2 y/o 3 meses.

En el presente trabajo se ha decidido definir 3 valores para d a la hora de calcular la

variable respuesta. Así se obtienen 3 tipos de variable respuesta a predecir para cada tipo de alisado aplicado a los precios de cada empresa. Estos 3 tipos hacen referencia a los distintos momentos en el futuro para los que se hace la predicción. En este caso, pues, se van a realizar predicciones sobre la dirección que tomará el precio a 1, 2 y 3 meses vista. De este modo se consigue una tabla 4x3 de combinaciones posibles para la definición del modelo entre el periodo de alisado exponencial y el número de días en el futuro sobre el que se quiere predecir. Teniendo en cuenta que en la base de datos sólo se tienen datos de los días laborables, se ha decidido establecer el valor de d acorde con la tabla siguiente para aproximar la predicción extactamente a 1, 2 y 3 meses:

Prediccion	d
1 mes	20
2 meses	40
3 meses	60

Tabla 5.11: Periodos de predicción

Las tablas siguientes muestran la proporción de observaciones que representan cada una de las dos clases de la variable respuesta en las distintas empresas para los distintos tipos de alisado aplicados:

Coca Cola CO.

	Predicción 1 mes		Predicción 2 meses		Predicción 3 meses	
	r1.Down	r1.Up	r2.Down	r2.Up	r3.Down	r3.Up
EMA30	45.44301	54.55699	43.25759	56.74241	41.37343	58.62657
EMA60	43.64758	56.35242	40.69246	59.30754	38.91393	61.08607
EMA90	40.71536	59.28464	38.97612	61.02388	39.25254	60.74746
Alisado exponencial	45.06095	54.93905	42.76066	57.23934	40.46206	59.53794

Tabla 5.12: Proporción de la variable respuesta Coca-Cola CO.

Como se puede observar en la tabla anterior la variable respuesta está relativamente bien balanceada en la mayoría de los casos. Cuanto más lisos són los datos, es decir, mayor el periodo de cálculo de las medias móviles exponenciales, mayor es la proporción de días en los que el precio al cabo de 1, 2 y 3 meses ha sido superior (Up). A su vez, para un mismo tipo de alisado sobre los datos, se observa que cuanto más lejos se predice la dirección del precio de cierre, mayor es el porcentage de días en los que el precio de cierre ha sido más elevado. En este caso, esto significa que si el *stock* presenta una tendencia creciente, como es el caso para esta empresa, el hecho de predecir su dirección en un futuro más lejano hace que se la variable respuesta se desbalancee a favor de los días con dirección Up.

Apple Inc.

	Predicción 1 mes		Predicción 2 meses		Predicción 3 meses	
	r1.Down	r1.Up	r2.Down	r2.Up	r3.Down	r3.Up
EMA30	34.08754	65.91246	31.83266	68.16734	30.30497	69.69503
EMA60	31.92169	68.07831	30.26288	69.73712	29.62009	70.37991
EMA90	28.82844	71.17156	28.26414	71.73586	27.45733	72.54267
Alisado exponencial	33.18621	66.81379	31.21570	68.78430	29.71598	70.28402

Tabla 5.13: Proporción de la variable respuesta Apple Inc.

En el caso de la empresa Apple Inc. la variable respuesta no está balanceada en ninguno de los casos. En la tabla 5.13 se puede observar que cuanto menor sea el nivel de alisado en los datos, correspondiente a la EMA de 30 días, más cerca estará la variable respuesta de estar balanceada (34% Down - 66% Up). Esto significa que cuanto más alisados sean los datos y más elevada sea d en la fórmula 2 más desbalanceada estará la variable respuesta para esta empresa, alcanzado un máximo de 27.3% Down - 72.7% Up. Esta proporción tan desbalanceada se explica por la fuerte tendencia creciente de los precios de cierre de AAPL dentro del período de estudio.

American Express CO.

	Predicción 1 mes		Predicción 2 meses		Predicción 3 meses	
	r1.Down	r1.Up	r2.Down	r2.Up	r3.Down	r3.Up
EMA30	41.48869	58.51131	42.25950	57.74050	38.68629	61.31371
EMA60	39.88083	60.11917	38.85446	61.14554	36.35973	63.64027
EMA90	36.45320	63.54680	36.45945	63.54055	34.13264	65.86736
Alisado exponencial	41.27785	58.72215	41.40988	58.59012	38.53328	61.46672

Tabla 5.14: Proporción de la variable respuesta American Express CO.

En el caso de la empresa American Express CO. la variable respuesta está en general ligeramente desbalanceada y presenta el mismo patrón de crecimiento del desbalanceo que las empresas Coca Cola CO. y Apple Inc, alcanzando el máximo balanceo entre las clases de la variable respuesta con la combinación de alisado con EMA 30 días y predicción a un mes vista (41% Down - 59% Up).

Wells Fargo & CO.

	Predicción 1 mes		Predicción 2 meses		Predicción 3 meses	
	r1.Down	r1.Up	r2.Down	r2.Up	r3.Down	r3.Up
EMA30	40.85430	59.14570	40.34827	59.65173	41.82128	58.17872
EMA60	39.66801	60.33199	39.53836	60.46164	39.83687	60.16313
EMA90	38.63782	61.36218	38.00817	61.99183	37.15705	62.84295
Alisado exponencial	40.87852	59.12148	40.12241	59.87759	41.28868	58.71132

Tabla 5.15: Proporción de la variable respuesta Wells Fargo and CO.

En el caso de la empresa Wells Fargo & CO. la variable respuesta está, en general, ligeramente desbalanceada y presenta el mismo patrón de crecimiento del desbalanceo que las empresas Coca Cola CO., Apple Inc y American Express CO.. La variable respuesta calculada para este *stock* presenta un máximo balanceo con la combinación de alisado con EMA 30 días y predicción a un mes vista (40.8% Down - 59.2% Up). El caso de esta empresa es distinto al de las anteriores en el hecho de que el balanceo entre las clases de la variable respuesta aumenta conforme se predice más en futuro para el caso de alisado exponencial de los datos. Cuando se analiza este tipo de alisado, la variable respuesta está más balanceada cuando se predice a 3 meses vista que cuando se predice a 1 mes vista.

Variables predictoras / explicativas

En el presente trabajo se utilizan distintos indicadores técnicos como variables predictoras de los modelos que se construyen posteriormente. Los indicadores técnicos son estadísticos que se calculan a partir de los precios OHLC que presenta un *stock*. Muchos managers de inversión e inversores en el mercado de *stocks* aceptan y utilizan generalmente ciertos indicadores técnicos como señales de las tendencias futuras del mercado (véase Kim, 2003). Algunos indicadores técnicos son efectivos en mercados con tendencia y otros rinden mejor en mercados no cíclicos y sin tendencia (véase Ray Tsaih, 1998). En mucha de la literatura revisada en torno a la previsión del precio de los *stocks* con indicadores técnicos queda probada la utilidad de éstos como variables predictoras en modelos de predicción tanto del precio como de la dirección de movimiento de los mismos (An-Sing Chena, 2003). En el presente trabajo se han seleccionado distintos indicadores técnicos para utilizarlos como variables predictoras a partir de la revisión de artículos de expertos en la materia (Kim, 2003, Han (2012), Yakup Kara (2011), C.L. Huang (2009), Manish Kumar (2006), Y. Nakamori (2005)). Seguidamente se detallan los indicadores técnicos utilizados y las fórmulas necesarias para calcularlos.

Para facilitar el cálculo de los siguientes indicadores se ha creado una función de `R` llamada `feature_extraction_finance`. Esta función recibe un `data.frame` con los precios OHLC y el volúmen de un *stock* y devuelve otro `data.frame` en el cual cada observación representa una fecha y cada columna una variable, siendo la primera de ellas la variable respuesta.

Aroon

Aroon es un indicador que puede descubrir el inicio de tendencias. Consiste en tres índices entre los cuales se escogen los dos primeros para utilizarlos como variables predictoras. Éstos están definidos en las fórmulas TI.1 y TI.2.

$$AroonUp = 100 * (\frac{n - PeriodSinceHighestHigh}{n}) \qquad (IT.1)$$

$$AroonDown = 100 * (\frac{n - PeriodSinceLowestLow}{n}) \qquad (IT.2)$$

Media móbil simple 10 días

La media móbil simple, SMA por sus siglas en inglés, calcula la media de los precios en un período de tiempo.

$$SMA_{10} = \frac{Close_t + Close_{t-1} + ... + Close_{t-10}}{10} \qquad (IT.3)$$

Momentum

Este indicador técnico mide la cantidad que el precio de un *stock* ha variado respecto d días en el pasado. En el presente trabajo se utilizan como variables predictoras los indicadores Momentum con $d = 1, 2, 3, 4, 5, 10$ y 15 calculados a partir de la fórmula (IT.4).

$$Mom_d = Close_t - Close_{t-d} \qquad (IT.4)$$

Rate of Change

El ratio de cambio, ROC por sus siglas en inglés, calcula la variación de un precio respecto d días en el pasado expresado en porcentage (véase ROC en Ulrich, 2018). En este trabajo se calcula el ratio de cambio respecto $d = 1$ y 9 días en el pasado utilizando la fórmula (IT.5).

$$ROC_d = (\frac{Close_t - Close_{t-d}}{Close_{t-d}}) * 100 \qquad (IT.5)$$

Stochastic oscillator: fast %K, fast %D & slow %D

El indicador técnico llamado Stochastic oscillator relaciona la localización del precio de cierre de cada día en relación con el rango de valores que ha tomado en los últimos d días. En el presente trabajo se utilizan como variables predictoras los estadísticos fast %K, fast %D y slow %D. El hecho de que el indicador sea llamado *fast* hace referencia a que los precios con los que se han calculado esos indicadores no han sido alisados. De manera contraria, el hecho de que el indicador sea llamado *slow* se refiere a que se ha calculado a partir de precios alisados. En este caso, al calcular los distintos indicadores a partir de precios ya alisados, el hecho de que un indicador sea *slow* significa que se ha calculado a partir de precios doblemente alisados. Los períodos utilizados para el cálculo de estos indicadores técnicos son $dFastK = 14, nFastD = 3, nSlowD = 3$, correspondientes a los valores por defecto de la función `stoch` del paquete `TTR`. Para el cálculo de estos indicadores se han utilizado las fórmulas siguientes:

$$fast\ K = (\frac{Close_t \quad - LowestLow_{t-d}}{HighestHigh_{t-d} - LowestLow_{t-d}}) * 100 \qquad (IT.6)$$

Dónde $LowestLow_{t-d}$ y $HighestHigh_{t-d}$ hacen referencia al precio de cierre mínimo más pequeño y al máximo más grande de los últimos $t - d$ días.

$$fast\ D = MovingAverage(fastK) = \frac{\sum_{i=0}^{n-1} fastK_{t-i}}{n} \qquad (IT.7)$$

$$slow\ D = MovingAverage(slowK) = \frac{\sum_{i=0}^{n-1} slowK_{t-i}}{n} \qquad (IT.8)$$

Stochastic Momentum Index

Este indicador técnico calcula el valor relativo del precio de cierre respecto el punto medio del rango de precios de los pasados d. Es parecido al oscilador estocástico pero con respecto al punto medio del rango de precios y no respecto al rango total. En otras palabras mide cuán cerca está el precio de cierre respecto el centro del rango de precios de los pasados d días. El cálculo de este indicador se desarrolla en las fórmulas siguientes:

$$cm = Close_t - \frac{(HighestHigh - LowestLow)}{2}$$

$$hl = HighestHigh - LowestLow$$

$$cmMA = EMA(EMA(cm))$$

$$hlMA = EMA(EMA(hl))$$

$$SMI = 100 * \frac{cmMA}{\frac{hlMA}{2}} \qquad (IT.9)$$

En la presente disertación se utilizado como característica de los datos el estadístico calculado con la fórmula (IT.9). Los períodos utilizados para el cálculo de las medias móbiles exponenciales son los valores por defecto (véase SMI en Ulrich, 2018).

Relative Strenght Index

El índice de fuerza relativa, llamado RSI por sus siglas en inglés, es un indicador que pretende medir la fuerza del *stock* en el movimiento que presenta actualmente. Calcula la ratio de los recientes precios crecientes respecto el movimiento absoluto de los precios (véase RSI en Ulrich, 2018). Se ha calculado a partir de la fórmula siguiente:

$$RSI = 100 - \frac{100}{1 + RS} \qquad (IT.10)$$

$$RS = \frac{ganancias\ medias}{prdidas\ medias} = \frac{\frac{\sum_{i=0}^{n-1} Up_{t-i}}{n}}{\frac{\sum_{i=0}^{n-1} Dw_{t-i}}{n}}$$

Dónde Up_t representa el cambio de los precios al alza y Dw_t el cambio de los precios a la baja en el momento t

Indicador Williams Accumulation/Distribution

Este indicador técnico pretende identificar la tendencia del mercado y medir la presión del mismo. Es llamado Williams AD por sus siglas en inglés y se calcula utilizando las fórmulas (IT.11), (IT.12) y (IT.13) (véase williamsAD en Ulrich, 2018):

Si $Close_t > Close_{t-1}$ *entonces*

$$AD_t = AD_{t-1} + Close_t - min(Low_t, Close_{t-1}) \qquad (IT.11)$$

Si $C_t < C_{t-1}$ *entonces*

$$AD_t = AD_{t-1} + max(High_t, Close_{t-1} - Close_t) \qquad (IT.12)$$

Si $C_t = C_{t-1}$ *entonces*

$$AD_t = AD_{t-1} \qquad (IT.13)$$

Indicador Williams Percentage Range

Este indicador técnico se calcula de un modo parecido al indicador estocástico fast %K. Es un indicador *momentum* que mide niveles de sobreventa y sobrecompre. Se ha calculado a partir de la fórmula (IT.14).

$$WPR = \frac{HighestHigh_n - Close_t}{HighestHigh_n - LowestLow_n} * 100 \qquad (IT.14)$$

Indicador Moving Average convergence divergence

Este indicador es más conocido por sus siglas en inglés: MACD. Este indicador es un oscilador que calcula la diferencia entre dos medial móbiles exponenciales, una "rápida" (n=12) y un "lenta" (n=26), las cuales pueden reflejar tendencias en el movimiento de los precios (véase MACD en Ulrich, 2018). Ha sido calculado a partir de la fórmula (IT.15)

MACD = EMA(stockPrices,12) − EMA(stockPrices,26) (IT.15)

Indicador Commodity Channel Index

Este indicador es comúnmente conocido como CCI por sus siglas en inglés. Se utiliza para descubrir los principios y finales de tendencias en *securities* (véase Lambert, 1980). Se calcula a partir de las fórmulas siguientes. El valor utilizado como variable predictora en este trabajo es el calculado con la fórmula (IT.16)

$$TP_t = \frac{High_t + Low_t + Close_t}{3}$$
$$MATP_t = MovingAverage(TP, n)$$

$$MDTP = \frac{\sum_{i=1}^{n} |TP_{t-i+1} - MATP_t|}{n}$$

$$CCI = \frac{TP_t - MATP_t}{0.015 MDTP_t} \qquad (IT.16)$$

Indicador On Balance Volume

Este indicador se llama OBV por sus siglas en inglés y mide el flujo de dinero dentro y fuera de un *stock*. Es una medida del flujo monetario que presenta un *stock*. Su cálculo viene dado por las fórmulas (IT.17), (IT.18) y (IT.19):

Si $Close_t > Close_{t-1}$ *entonces*

$$OBV_t = OBV_{t-1} + Volume_t \qquad (IT.17)$$

De otro modo si $Close_t < Close_{t-1}$ *entonces*

$$OBV_t = OBV_{t-1} - Volume_t \qquad (IT.18)$$

De otro modo

$$OBV_t = OBV_{t-1} \qquad (IT.19)$$

Indicador Average True Range

El indicador rango medio verdadero, ATR por sus siglas en inglés, es un grupo de estadísticos que estima la volatilidad de una serie temporal (Wilder, 1978). Su cálculo viene dado por las fórmulas siguientes:

$$TrueHigh = max(High_t, Close_{t-1})$$
$$TrueLow = min(Low_t, Close_{t-1})$$

TR_t= TrueHigh_t − TrueLow_t (IT.20)

$$ATR = \frac{TR_{t-1} * (n \ menos \ 1) + TR_t}{n} \qquad (IT.21)$$

En el presente trabajo se utilizan como variables predictoras las fórmulas del verdadero rango y el verdadero rango medio definidas en las ecuaciones (IT.20) y (IT.21). Además se

utiliza también una variante del verdadero rango calculada en la fórmula (IT.22)

TR2_t = (Close_t − TrueLow_t)/(TrueHigh_t − TrueLow_t) (IT.22)

Indicador Trend Detection Index

Este índice de detección de tendencias se llama TDI por sus siglas en inglés. Se utiliza para descubrir el inicio y el final de tendencias móbiles (véase TDI en Ulrich, 2018). Su cálculo viene dado por las ecuaciones siguientes:

Mom = precio − precio_{periodo}

$$MomAbs = |Mom|$$

$$DI = \sum_{i=1}^{n} Mom_i \qquad (IT.23)$$

$$DIAbs = |DI|$$

$$DIAbsSum = \sum_{i=1}^{n} DIAbs$$

$$DIAbsSum2 = \sum_{i=1}^{2n} DIAbs$$

TDI = DIAbs−(DIAbsSum2−DIAbsSum) (IT.24)

En la presente disertación se utilizan como variables predictoras los indicadores DI y TDI definidos en las ecuaciones (IT.23) y (IT.24)

Indicador Welles Wilder's Directional Movement Index

Este indicador se conoce generalmente como ADX. ADX consiste en 4 indicadores llamados Índice Direccional Poisitvo (DIp), Índice Direccional Negativo (DIn), Índice Direccional (DI) y el Índice direccional Medio (ADX) (véase Wilder, 1978).

HiDiff = High_{t-1} − High_t

LoDiff = Low_{t} − Low_{t-1}

$Si\ HiDiff < 0\ y\ LoDiff < 0\ \ o\ \ HiDiff = LoDiff\ entonces$

$$DIp = 0$$

$$DIn = 0$$

$Si\ HiDiff > LoDiff\ \ entonces$

$$DIp = HiDiff$$

$$DIn = 0$$

$Si\ HiDiff < LoDiff\ \ entonces$

$$DIp = 0$$

$$DIn = LoDiff$$

$$DX = \frac{DIp\ DIn}{DIp + DIn} \qquad (IT.25)$$

$$ADX = \frac{ADX_{t-1} * (n\ menos\ 1) + DX_t}{n} \qquad (IT.26)$$

En el presente trabajo se utilizan las ecuaciones (IT.25) y (IT.26) como variables predictoras en el modelo. Por añadidura también se utiliza como variable predictora la ratio dada en la ecuación (IT.27)

$$PN_{ratio} = \frac{DI_p}{DI_n} \qquad (IT.27)$$

Indicador Bollinger Bands

Este indicador llamado Bandas de Bollinger, más conocido por sus siglas en inglés BB, es utilizado para medir la volatilidad de un *stock* y compararla con el nivel para un período de tiempo (véase BBands en Ulrich, 2018). Se ha calculado a partir de la fórmula siguiente (IT.28).

$$TP_t = \frac{High_t + Low_t + Close_t}{3}$$

$$BandWidth_t = 2 * F * (TP_t) \qquad (IT.28)$$

En este trabajo se utiliza la ecuación IT.28 como variable. Las bandas superior e inferior están a F desviaciones típicas por encima y por debajo de la banda intermedia. En este caso $F = 2$.

En todos los casos:

- $Close_t$ es el precio de cierre del stock en el momento t
- Low_t es el precio mínimo del stock en el momento t
- $High_t$ es el precio máximo del stock en el momento t

CUANDO HACEMOS EL ALISADO EXPONENCIAL DE LOS DATOS CON LA MEDIA EXPONENCIAL LO QUE ESTAMOS HACIENDO ES INTENTAR PREDECIR LA DIRECCIÓN QUE TOMARÁ ESA MEDIA AL CABO DE 1, 2, 3 MESES. ES DECIR ESTMOS MIRANDO DE PREDECIR SI LA MEDIA A 30, 60,90 DIAS ESTARÁ MÁS ALTA EN 1, 2, 3 MESES

CUANDO MIRAMOS INTENTAMOS PREDECIR QUE LA MEDIA EXP 30 DIAS ESTARÁ MÁS ALTA EN 30 DÍAS SIGNIFICA QUE INTENTAMOS PREDECIR SI EN UN MES HABRÁ UN INCREMENTO DE LA MEDIA DEL PRECIO DE CIERRE DE ESE MES AND SO ON

NO SE REALIZA NINGUNA NORMALIZACIÓN PARA EL RANDOM FOREST YA

QUE ESTE MODELO NO ES SENSIBLE A TRANSFORMACIONES MONÓTONAS DE LOS DATOS

CAPÍTULO 6

Modelización

6.1 Definiciones de los modelos

En este apartado se detallan los distintos modelos utilizados para intentar predecir la dirección de movimiento de los stocks. Todos los modelos que se definen y construyen en el presente trabajo son modelos de clasificación binaria al estar definida la variable respuesta como Up and Down.

Random Forest

Los árboles de decisión CART, llamados así por su nombre en inglés `Classification and Regression Trees`, son un tipo de modelos que se pueden utilizar para distintos tipos de aplicaciones de aprendizaje automático. En resumidas palabras, el método consiste en partir los datos a partir de un valor de cierta variable. Cada nodo padre genera 2 nodos hijos al tratarse de un problema de clasificación donde la variable respuesta tiene 2 clases. Esta partición se hace a partir de un criterio de impureza de los datos de manera que los nodos finales, llamados hojas, tengan la mayor pureza posible. Sin embargo los árboles que se hacen crecer de una manera muy profunda, es decir árboles muy grandes en cuanto al número de *split* y la profundidad que cogen, para aprender patrones altamente irregulares tienden a sobreajustar los datos de entrenamiento (problema conocido en inglés como *overfitting*). Un ligero ruido en los datos puede causar que el árbol crezca de una manera completamente diferente (Luckyson Khaidem, 2016, p. 7). Esto se debe al hecho de que los árboles de decisión tienen poco sesgo pero una alta varianza, al hacer pocas asumpciones sobre la variable respuesta (sesgo) pero altamente susceptibles a las variables predictoras (varianza). En otras palabras, un árbol de decisión casi no hace asumpciones sobre la variable objetivo (sesgo pequeño) pero es altamente susceptible a variaciones de los datos que se utilizan como input (alta varianza). Seguidamente se muestra una ejemplo sencillo de cómo un árbol de decisión luce. Esta imagen representa el ejemplo en el que se quiere predecir el género de una persona a partir de se altura y su peso.

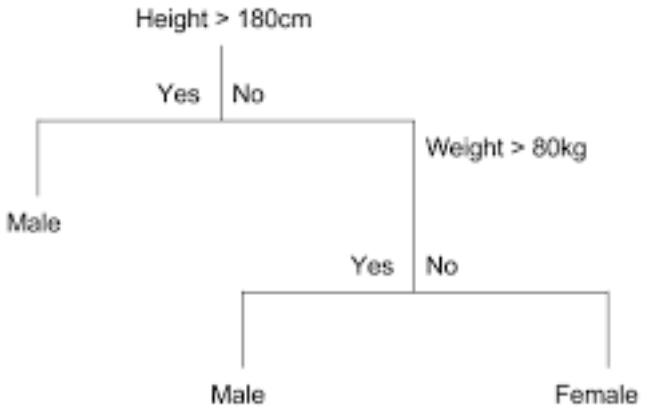


Figura 6.1: Ejemplo de árbol de decisión

En este punto es cuando aparece el modelo de aprendizaje automático llamado `Random Forest`. Este tipo de modelos superan el problema explicado en el párrafo anterior entrenando múltiples árboles de decisión en un subespacio del espacio formado por las variables predictoras/explicativas, asumiendo como coste un ligero incremento del sesgo. Esto significa que ninguno de los árboles del bosque es entrenado con la totalidad de los datos de entrenamiento. Los datos son recursivamente partidos en particiones, de manera que en un nodo particular la partición se elabora haciendo una "pregunta" a una de las variables. Por ejemplo, una partición podría estar hecha "preguntándole" a la variable *Rate of Change 1 day* cuántos datos tienen un valor superior/inferior a un cierto valor X. La elección del criterio de partición de los datos se basa en alguna medida de impureza tales como la Entropía de Shannon o la medida de impureza de Gini. En el presente trabajo se utiliza la función `randomForest` implementada en el paquete de R con el mismo nombre. Esta función utiliza como medida de impureza a partir de la cual se particionan los datos el índice de impureza de Gini (Liaw & Wiener, 2002). Este índice se utiliza como la función para medir la calidad de cada partición en cada nodo. La impureza de Gini en el nodo N se calcula a partir de la fórmula siguiente:

$$g(N) = \sum_{i \neq j} P(w_i)P(w_j)$$

Dónde $P(w_i)$ es la proporción de casos donde la variable respuesta toma la clase i. $P(w_j)$ entonces es la propoción de casos en los cuales la variable respuesta toma la clase j.

La manera heurística para escoger la mejor decisión de particion en un nodo específico

se basa en el hecho de conseguir la mayor reducción posible de impureza. Es decir, la mejor partición posible en un determinado nodo viene definida por la mayor ganancia de información (variable que mejor particiona los datos / inclye más observaciones en cada partición) o por la mayor reducción de impureza. La ganancia de información que genera una determinada partición se puede calcular con la fórmula siguiente:

$$\triangle I(N) = I(N) - P_L * I(N_L) - P_R * I(N_L)$$

Dónde $I(N)$ es la medida de impureza (ya sea la impureza de Gini o la entropía de Shannon) de un nodo N. P_L es la proporción de casos que en el nodo padre N van a parar al hijo **izquierdo**. De un modo similar, P_R representa la proporción de casos en el nodo padre N que se van a parar al nodo hijo **derecho** después de realizar la partición. N_L y N_R son los nodos hijos izquierdo y derecho, respectivamente.

Este tipo de modelos de aprendizaje automático son conocidos como modelos *ensemble*. En el núcleo de estos modelos está el *Bootstrap aggregating*, mayormente conocido como *bagging*. Esto significa que la predicción final se calcula como una media de la solución obtenida con cada árbol construido sobre cada remuestra generada con la técnica no paramétrica del *bootstrap*. En otras palabras: utilizando *bootstrap* se calculan remuestras de los datos con las cuales se contruye un árbol. Dentro de cada árbol calculado sobre cada remuestra *bootstrap* cada nodo es partido utilizando la mejor variable dentro de la muestra de variables escogidas aleatoriamente en cada nodo. Al final, la predicción del modelo es una media de los valores obtenidos con todos estos árboles calculados sobre las distintas remuestras *bootstrap* (véase Liaw & Wiener, 2002, p. 18 y Dietterich (n.d.)). El método *bagging* mejora la estabilidad y la precisión de los algoritmos de aprendizaje. Al mismo tiempo reduce la varianza y el sobreajuste, los cuales son un problema relativamente común al construir árboles de decisión CART (véase Luckyson Khaidem, 2016, p. 8 para un resumen del algoritmo escrito en pseudocódigo).

Regresión logística

Deep Neural Network Classifier

6.2 Métricas de rendimiento

Los modelos construidos están pensados para ser de ayuda a la hora de tomar una decisión de compra o venta de un *stock*. Si la predicción es +1 (Up) se espera que el precio al cabo de $d=30, 60 y 90 $ días sea superior al actual y entonces la acción lógica del inversor sería **comprar**. Al contrario ocurriría si la predicción fuera -1, lo cual significaría que se espera que el precio al cabo de d días sea inferior. En este caso la decisión razonable a tomar por parte del inversor sería **vender**. Una predicción errónea puede causar grandes perdidas de dinero para el inversor. Por lo tanto, se deben definir métricas que ayuden a evaluar la potencia

predictiva de los modelos construidos. Las métricas que se utilizan en el presente trabajo para evaluar la robustez del clasificador binario Random Forest que se crea a continuación son las siguientes:

Accuracy

Esta métrica hace referencia a la proporción de casos clasificados correctamente entre el total de casos con los que se prueba el modelo. Se calcula a partir de la fórmula siguiente:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \qquad (M.1)$$

Recall/Sensitivity

Esta métrica mide la habilidad de un modelo clasificador de identificar correctamente los casos positivos.

$$Sensitivity = \frac{tp}{tp + fn} \qquad (M.2)$$

Specificity Esta métrica mide la habilidad de un modelo clasificador de identificar correctamente los casos negativos.

$$Specificity = \frac{tn}{tn + fp} \qquad (M.3)$$

Dónde,

tp = número de verdaderos positivos ≡ número de veces en las que el caso era positivo y el modelo predice positivo tn = número de verdaderos negativos ≡ número de veces en las que el caso era negativo y el modelo predice negativo fp = número de falsos positivos ≡ número de veces en las que el caso era negativo y el modelo predice positivo fn = número de falsos negativos ≡ número de veces en las que el caso era positivo y el modelo predice negativo

6.3 Experimentos

Partición muestras de entrenamiento, validación y test

Una vez calculada la variable respuesta y las variablex explicativas/predictoras se procede a construir los modelos predictivos. En primer lugar se particionan los conjuntos de datos en las muestras de entrenamiento, validación y test. En la muestra de entrenamiento se encuentran los datos con los que se entrarán los modelos, es decir, son los datos a partir de los cuales los modelos van observar el fenómeno de estudio y van a aprender sobre él. La muestra de validación se utiliza para obtener una optimización de los hyper parámetros. En inglés este proceso se llama *hyper-parameter fine tunning*. Finalmente, la muestra test la forman los datos sobre los cuales se va a evaluar el rendimiento de los modelos entrenados con los datos *train + validation* después de tunear los hyper parámetros.

Las series temporales de precios necesitan un tratamiento especial al momento de hacer las particiones entre muestras de entrenamiento, validación y test ya que se debe mantener la estructura temporal inherente en los datos.

- Muestra de entrenamiento: 2000-2015
- Muestra de validación: 2016
- Muestra de test: 2017-2018

Los modelos que se construyen requieren una muestra de entrenamiento considerable al tener que captar largas tendencias en los precios. Además los hyper parámetros que se deben optimizar, en el caso del Random Forest, no son demasiados y por eso se decide utilizar 17 años de entrenamiento y sólo 6 meses para la muestra de validación y test.

PONER AQUÍ EL NÚMERO TOTAL DE EXPERIMENTOS: CUANTOS MODELOS SE HACEN DE CADA TIPO, CONTANDO PARAMETER FINE TUNING Y TEST PERFORMANCE

Modelización y resultados

Random Forest

Para realizar los experimentos utilizando este modelo descrito en el apartado 6.1 del presente trabajo se optimiza el parámetro `mtry` (véase Philipp Probst & Boulesteix, 2018). Acorde con la documentación de la función `randomForest` que se puede consultar en (Liaw & Wiener, 2002), el parámetro `mtry` controla el número de variables aleatóriamente muestreadas que son candidatas en cada *split* del árbol. Es decir, es el número de variables a tener en cuenta en cada *split* a partir de las cuales se selecciona la que mejor particiona los datos. Esta optimización se hace sobre la muestra de validación utilizando como métrica el % de casos mal clasificados. Esto significa que se selecciona el valor de `mtry` que propocione un

menor porcentage de casos mal clasificados (1-accuracy).

Seguidamente se muestra una tabla resumen con el valor óptimo del hyper parámtero `mtry` para cada tipo de alisado de cada empresa. Además se incluye en la tabla el valor de *accuracy* obtenido con cada modelo sobre la muestra de validación utilizando el parámetro óptimo `mtry` para cada tipo de alisado.

	Predicción 1 mes		Predicción 2 meses		Predicción 3 meses	
	mtry1m	accuracy1m	mtry2m	accuracy2m	mtry3m	accuracy3m
EMA30	13	63.09524	2	54.76190	2	60.71429
EMA60	2	63.49206	2	56.34921	2	49.60317
EMA90	2	58.33333	4	75.00000	2	67.46032
Alisado exponencial	2	63.09524	16	66.26984	3	63.49206

Tabla 6.1: Coca Cola CO.: valores optimizados para mtry y accuracy obtenida

Como se puede observar para la empresa Coca-Cola CO. el mejor resultado, es decir la *accuracy* obtenida más alta, para la predicción a 1 mes se obtiene con el alisado correspondiente a una media móbil exponencial a 30 días con un valor de 63.5%. Para la predicción a 2 meses el mejor resultado es de 75% de accuracy y se obtiene alisando los datos con una EMA de 90 días (el mejor resultado global para la muestra de validación). Por lo que respecta a la predicción a 3 meses el mejor resultado se obtiene alisando los datos con una EMA 90 días con un valor de 67.46%. Para esta empresa no se aprecia un patrón de crecimiento de la *accuracy* sino más bien unos resultados repartidos más o menos uniformente entre los distintos experimentos realizados.

	Predicción 1 mes		Predicción 2 meses		Predicción 3 meses	
	mtry1m	accuracy1m	mtry2m	accuracy2m	mtry3m	accuracy3m
EMA30	29	79.36508	11	67.46032	14	66.66667
EMA60	26	80.15873	2	66.66667	23	72.22222
EMA90	28	87.30159	2	69.84127	18	73.80952
Alisado exponencial	27	78.96825	11	68.65079	11	63.09524

Tabla 6.2: Apple Inc.: valores optimizados para mtry y accuracy obtenida

Como se puede apreciar en la tabla 6.2, para la empresa Apple Inc. la situación es diferente. Se obtienen, en general, mejores resultados que en el caso de la empresa Coca Cola. Para la predicción a 1 mes el mejor resultado se obtiene con un alisado utilizando EMA de 90 días con una *accuracy* del 87.30%. Para la predicción hecha a 2 meses el mejor resultado es un 69.84% de accuracy obtenido con un alisado EMA 90 días. Para el caso en el que la predicción es a 3 meses el mejor resultado es de un 73.81% de *accuracy* con un alisado de EMA 90 días. Es decir, en el caso de la empresa Apple Inc el mejor resultado global se obtiene prediciendo si la media móbil exponencial a 90 días del precio de cierre estará más alta o no a 1 meses vista. En este caso los mejores resultados se obtienen con una predicción

hecha a 1 mes vista con los datos fuertemente alisados. El hecho de predecir si al cabo de un mes la media móbil exponencial del precio de cierre calculada con 90 días será más alta o no, es de ayuda para el inversor ya que significa que el precio de cierre al cabo de un mes será más alto. Esto se debe al hecho de que si la EMA 90 días será más alta al cabo de un mes significa que los precios de ese último mes han sido más altos.

	Predicción 1 mes		Predicción 2 meses		Predicción 3 meses	
	mtry1m	accuracy1m	mtry2m	accuracy2m	mtry3m	accuracy3m
EMA30	9	79.76190	2	53.96825	2	63.88889
EMA60	2	73.80952	2	54.36508	5	76.19048
EMA90	10	74.60317	31	57.93651	2	80.15873
Alisado exponencial	2	79.76190	3	59.92063	4	77.38095

Tabla 6.3: American Express CO.: valores optimizados para mtry y accuracy obtenida

Los resultados obtenidos para la empresa American Express CO. son globalmente mejores en los casos en los que la predicción está hecha para el siguiente mes y para al cabo de 3 meses. Para el caso en el que la predicción se hace a 1 mes, el mejor resultado obtenido es de 79.76% de *accuracy* en la muestra de validación y se corresponde a un alisado de los datos utilizando la fórmula del alisado exponencial. A su vez se obtiene el mismo resultado con los datos alisados mediante una media móbil exponencial de 30 días. Para el caso en el que la predicción se hace a 2 meses el mejor resultado que se obtiene es de 59.92% de *accuracy* en el caso de alisar los datos con la fórmula del alisado exponencial. Este resultado es ligeramente superior al que se obtendría con una predicción aleatória (50%). El resultado que se obtiene en el caso de la predicción hecha a 3 meses ya que en el mejor de los casos, cuando se alisan los datos con una media móbil exponencial de 90 días, es de 80.16% de *accuracy* (mejor resultado global). Esto significa que, para esta empresa, la mejor clasificación se obtiene prediciendo si la EMA 90 días será más alta o no al cabo de 3 meses. En esencia, este caso de predicción representa que se está analizando si el incremento de los precios ha hecho que la media móbil exponencial haya subido los últimos 3 meses.

	Predicción 1 mes		Predicción 2 meses		Predicción 3 meses	
	mtry1m	accuracy1m	mtry2m	accuracy2m	mtry3m	accuracy3m
EMA30	4	67.85714	2	50.00000	2	41.66667
EMA60	3	64.68254	2	50.39683	2	54.36508
EMA90	3	65.87302	2	55.15873	13	64.28571
Alisado exponencial	4	69.84127	3	58.33333	3	55.95238

Tabla 6.4: Wells Fargo and CO.: valores optimizados para mtry y accuracy obtenida

En el caso de Wells Fargo & CO. los resultados sobre la muestra de validación se pueden observar en la tabla 6.4. Son, en general, peores que los resultados obtenidos con la empresa Apple Inc. y American Express CO.. Para la predicción hecha a 1 mes el mejor resultado se obtiene utilizando la fórmula de alisado exponencial (2) con una *accuracy* de 69.84%

(mejor resultado global para esta empresa). Esto significa que sobre la muestra de validación aproximadamente el 70% de los casos queda bien clasificado. En el caso de la predicción a 2 meses el mejor resultado obtenido es de 58.33% de *accuracy* obtenido alisando los datos con la fórmula de alisado exponencial (2). El mejor resultadao cuando la predicción se hace a 3 meses vista es del 64.28% de *accuracy*.

Resultados sobre muestra test

Una vez optimizados los valores del hyper parámetro `mtry` se entrenan los modelos con las muestras de entrenamiento + validación con el objetivo de testar el modelo con los datos de test. En total se construyen 48 modelos: 4 por cada tipo de variable respuesta y, a su vez, 4 por cada empresa. En las tablas siguientes se muestran los valores de *accuracy*, *sensitivity* y *specificity* de los distintos modelos sobre los datos test.

	Predicción 1 mes			Predicción 2 meses			Predicción 3 meses		
	acc1m	sensi1m	speci1m	acc2m	sensi2m	speci2m	acc3m	sensi3m	speci3m
EMA30	41.16	95.21	17.61	45.12	94.17	27.86	58.05	98.17	44.88
EMA60	64.45	86.51	56.62	77.66	91.67	73.37	75.51	85.71	72.04
EMA90	84.20	92.56	81.39	65.08	93.28	55.26	69.84	91.53	61.92
Alisado exponencial	48.65	90.91	32.66	43.60	93.86	27.09	66.44	89.09	58.91

Tabla 6.5: Coca Cola CO.: Metricas de rendimiento sobre muestra test

	Predicción 1 mes			Predicción 2 meses			Predicción 3 meses		
	acc1m	sensi1m	speci1m	acc2m	sensi2m	speci2m	acc3m	sensi3m	speci3m
EMA30	67.15	60.16	69.69	51.41	42.05	53.62	18.37	100.00	9.09
EMA60	64.86	77.17	61.95	55.31	61.54	54.94	59.64	13.04	62.20
EMA90	74.22	63.27	75.46	63.99	27.27	65.83	19.50	100.00	16.86
Alisado exponencial	54.26	56.78	53.44	52.28	97.10	44.39	41.72	62.50	39.65

Tabla 6.6: Apple Inc.: Metricas de rendimiento sobre muestra test

Tabla 6.7: American Express CO.: Metricas de rendimiento sobre muestra test

Tabla 6.8: Wells and Fargo CO.: Metricas de rendimiento sobre muestra test

PONER LAS TABLAS CON LA MEDIA DE ACCURACY OBTENIDA ENTRE TODOS LOS ALISADOS. CREAR REGLA: SI COMO MINIMO 3 ALISADOS DICEN QUE SUBIRA, ES MAS ROBUSTO…

Importancia de las variables

```
##       ROC1       mom1        tdi         di       sma10        OBV
```

```
## 512.152044  98.939509  96.630552  88.196759  86.062215  83.901033
##        atr        ADX         wAD         DX        tr2         tr
##  81.766291  70.274946  67.530958  65.073874  51.855076  49.960171
##    BBwidth       macd        SMI    PNratio        RSI      mom15
##  49.893047  49.413250  49.311654  42.877846  41.831672  37.362612
##       ROC9       mom9       mom2       slowD       pctB        cci
##  34.654206  28.161904  27.327193  25.393657  22.006062  21.435058
##        WPR       mom5      fastK       mom3      fastD       mom4
##  20.790710  20.790427  20.708526  19.217105  18.165741  17.651639
##    aroonUp    aroonDn
##   8.278372   7.877230
```

Experimentos con los datos sin alisar

*poner aqui los precios reales para poder comparar… si se predice con la ema90 que subira, es verdad que en ese momento el precio estava mas alto que cuando se hizo la predicción?

CAPÍTULO 7

Resultados de los experiemtos

Keep it going.

7.2 Resultados y análisis

*poner para la prediccion a distintos meses cual es la empresa que mejor resultado da y el alisado etc

CAPÍTULO 8

Conclusiones

Well done.

CAPÍTULO 9

Bibliografía

An-Sing Chena, H. D., Mark T. Leungb. (2003). Application of neural networks to an emerging financial market: Forecasting and trading the taiwan stock index. *Computers & Operations Research, 30 (6) (2003)*, 901–923. Retrieved from https://ac.els-cdn.com/S0305054802000370/1-s2.0-S0305054802000370-main.pdf?_tid=a9ff1500-f141-436c-9298f0bf19d9c5c0&acdnat=1545913123_1b0ff6cdd7d647a8c5e9891ad3f6ea65

Andrew K. Rose, M. M. S. (2011). Cross-country causes and consequences of the 2008 crisis: Early warning. *Japan and the World Economy*.

C.L. Huang, C. T. (2009). A hybrid sofm-svr with a filter-based feature selection for stock market forecasting. *Expert Systems with Applications, 36 (2) (2009)*, 1529–1539. Retrieved from https://ac.els-cdn.com/S0957417407006069/1-s2.0-S0957417407006069-main.pdf?_tid=ffd2e07d-4100-4d76bc64-653d8ae68de7&acdnat=1545913369_ce2c3b4a4eb42518498e95713c23e5c3

DeLong, J. B. (2009). The financial crisis of 2007–2009: Understanding its causes, consequences—and its possible cures. *MTI-CSC Economics Speaker Series Lecture*.

Dietterich, T. G. (n.d.). Ensemble methods in machine learning. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.4718rep=rep1type=pdf

Gilliland, M. (2009). The coefficient of variation for assessing forecastability. *The Business Forecasting Deal*.

Han, Z. (2012). Data and text mining of financial markets using news and social media, 1–102. Retrieved from https://studentnet.cs.manchester.ac.uk/resources/library/thesis_abstracts/MSc12/FullText/Han-Zhichao-fulltext.pdf

Kim, K.-j. (2003). Financial time series forecasting using support vector machines. *Neurocomputing 55 (2003)*, 307–319. Retrieved from https://ac.els-cdn.com/S0925231203003722/1-s2.0-S0925231203003722-main.pdf?_tid=ec7d432b-e65c-434eb695-6d2035ff9829&acdnat=1545911976_ef14d01ea7992160caa0d8b4aa7c13aa

Lambert, D. (1980). Commodities(now called futures).

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22. Retrieved from https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf

Luckyson Khaidem, S. R. D., Snehanshu Saha. (2016). Predicting the direction of stock

market prices using random forest. *To Appear in Applied Mathematical Finance*, 21.

Manish Kumar, M. T. (2006). Forecasting stock index movement: A comparison of support vector machines and random forest. *Indian Institute of Capital Markets 9th Capital Markets Conference Paper*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=876544

Masoud, N. (2014). Predicting direction of stock prices indexmovement using artificial neural networks:The case of libyan financial market. *British Journal of Economics, Management & Trade*, 597–619. Retrieved from https://www.researchgate.net/publication 272758765_Predicting_Direction_of_Stock_Prices_Index_Movement_Using_Artificial_Neural_Networks_The_Cas

P.R. Lane, G. M.-F. (2011). The cross-country incidence of the global crisis. *IMF Economic Review*, *59*, 77–110.

Philipp Probst, M. W., & Boulesteix, A.-L. (2018). Hyperparameters and tuning strategies for random forest. Retrieved from https://arxiv.org/pdf/1804.03515.pdf

Ray Tsaih, C. C. L., Yenshan Hsu. (1998). Forecasting s&P 500 stock index futures with a hybrid ai system. *Decision Support Systems, 23 (1998)*, 161–174. Retrieved from https://ac.els-cdn.com/S0167923698000281/1-s2.0-S0167923698000281-main.pdf_tid=5c0465f0-1321-4ffd-920b-4ff50ab6fd3b&acdnat=1545912698_86722ea4bbc2cf26a7590b8bdfeab92d

Ulrich, J. (2018). *TTR: Technical trading rules*. Retrieved from https://CRAN.R-project.org/package=TTR

Urban Jermann, V. Q. (2003). Stock market boom and the productivity gains of the 1990s.

Warren buffett : Latest portfolio. (n.d.). http://warrenbuffettstockportfolio.com/.

Wilder, J. (1978). New concepts in technical trading systems. *Trend Research Greensboro, North Carolina*.

Y. Nakamori, S. W., W. Huang. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research, 32 (10)*, 2513–2522.

Yakup Kara, Ö. K. B., Melek Acar Boyacioglu. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert Systems with Applications*, 5311–5319. Retrieved from https://ac.els-cdn.com/S0957417410011711/1-s2.0-S0957417410011711-main.pdf?_tid=e5e12593-4a56-4db0-9d5f ff9784196726&acdnat=1545074291_d05308f34977383caaff6ae009a6acf5