

Predicción de Series Temporales Financieras con Machine Learning

Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus Aplicaciones - Edición 2024

Mentor: Emmanuel Tassone

Integrantes:

- Gonzalez, Juan Cruz
- Fullana Jornet, Marcelo Fernando
- Herrador, Emanuel Nicolás
- Itovich, Griselda

FAMAF

Facultad de Matemática,
Astronomía, Física y
Computación



UNC

Universidad
Nacional
de Córdoba



CCAD

CÓRDOBA
CLUSTER



Ministerio de
CIENCIA Y
TECNOLOGÍA



Parte 1

Generalidades del proyecto y del
conjunto de datos

- Objetivo
- Dataset de base
- Preprocesamiento

Objetivo

Realizar y comparar diferentes modelos de predicciones con distintas técnicas de Machine Learning, con el fin de anticipar los movimientos de las acciones de Tesla



Dataset

date	open	high	low	close	volume	change_percent	avg_vol_20d
07/27/2010	20.91	21.18	20.2599	20.55	619675	-1.91	74326218.5
07/28/2010	20.55	20.9001	20.5101	20.7201	467183	0.82	60589147.1
07/29/2010	20.7699	20.88	20.0001	20.3499	615910	-1.78	48155285.4
07/30/2010	20.1999	20.4399	19.5501	19.9401	426830	-2.02	42303010.9
08/02/2010	20.4999	20.97	20.3331	20.9199	719145	4.92	38986014.6
08/03/2010	21	21.9501	20.82	21.9501	1231022	4.92	34749810
08/04/2010	21.9501	22.1799	20.85	21.2601	920755	-3.14	30246691.7
08/05/2010	21.54	21.5499	20.0499	20.4501	796379	-3.81	25054321.9
08/06/2010	20.1	20.16	19.5201	19.59	742138	-4.2	22566971.1
08/09/2010	19.8999	19.98	19.4499	19.5999	812655	0.05	21523785.6
08/10/2010	19.65	19.65	18.8199	19.0299	1281285	-2.91	20474705.1
08/11/2010	18.69	18.8799	17.85	17.9001	797649	-5.94	17925860.2
08/12/2010	17.7999	17.9001	17.3901	17.6001	691281	-1.68	15635348.3
08/13/2010	18.18	18.45	17.6604	18.3201	634513	4.09	14145325.7
08/16/2010	18.45	18.8001	18.2616	18.78	486211	2.51	12645117.3

Trabajamos con los precios de cierre (close) que cubren los registros diarios del mercado de Tesla en el periodo 1 enero 2020-15 de abril 2024 (total de 1567 casos/registros).

Precio de cierre (close)

Precio de cierre de Tesla



Preprocesamiento datos

-**Interpolación** de datos faltantes (feriados y fines de semana)

-Variables predictoras (Lag):

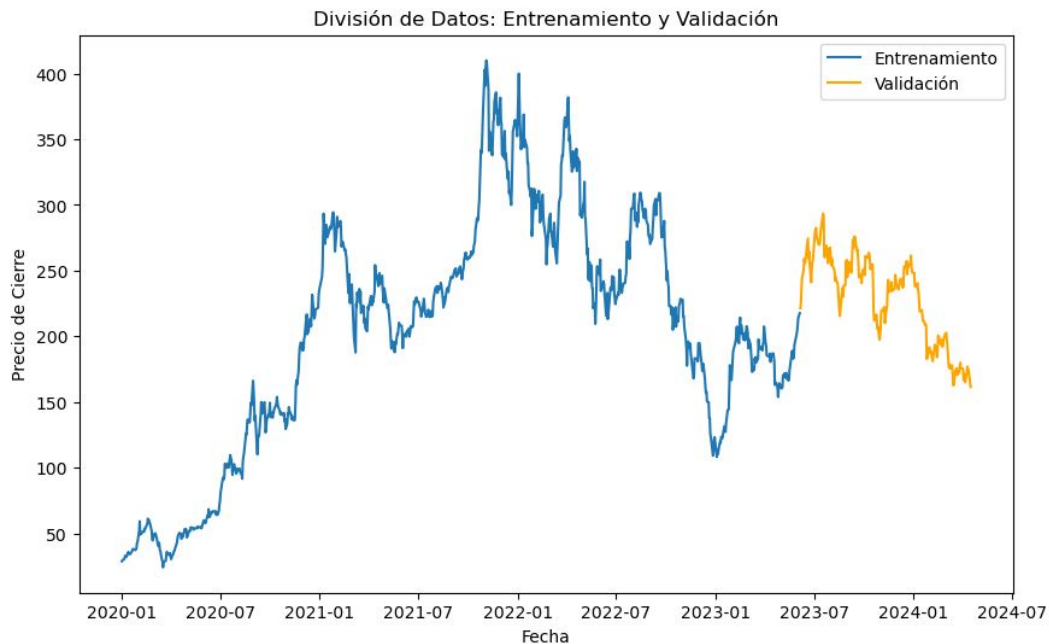
las variables Lag es la forma clásica para convertir problemas de series temporales en problemas de aprendizaje supervisado. Consiste en utilizar los valores pasado de las series para predecir el futuro. Utilizamos 30 variables lags, lo que significa que los modelos utilizan los valores de precio de cierre de los últimos 30 días para predecir el valor del día siguiente.

Lag1	Original
NaN	150.3
150.3	152.7
152.7	149.8



Entrenamiento y validación

A diferencia de los conjuntos de datos tradicionales, donde los datos de entrenamiento y validación se pueden seleccionar al azar, en series temporales hay que **mantener el orden cronológico**. Esto se debe a que las series temporales dependen de la secuencia de los eventos en el tiempo, y una selección aleatoria rompería esta estructura.



Parte 2

Pronósticos efectuados
con modelos de Machine
Learning

- Regresión Lineal
- Random Forest Regressor
- Support Vector Regressor
- ARIMA
- Redes neuronales:
 - Perceptron
 - LSTM

Modelos de predicción:

Regresión lineal (RL) - Random Forest Regressor (RFR) - Support Vector Regressor (SVR)

Periodo seleccionado : 01/01/20 al 15/04/24 (completo en rango de días y sin nans)

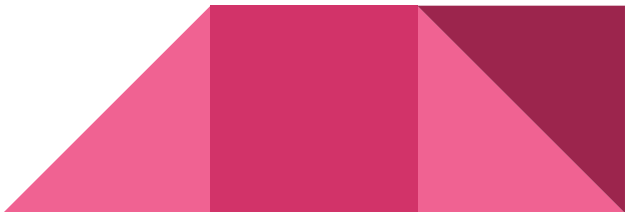
Columnas para predecir: 'close' y los 30 días previos a cada registro (30 columnas lagueadas)

Entrenamiento y testeo : 70% y 30%

Metodología: Se buscaron los mejores resultados con cada técnica por defecto y mediante una grilla de parámetros.

Ranking por rendimiento general en testeo (según métricas MAE y RMSE):

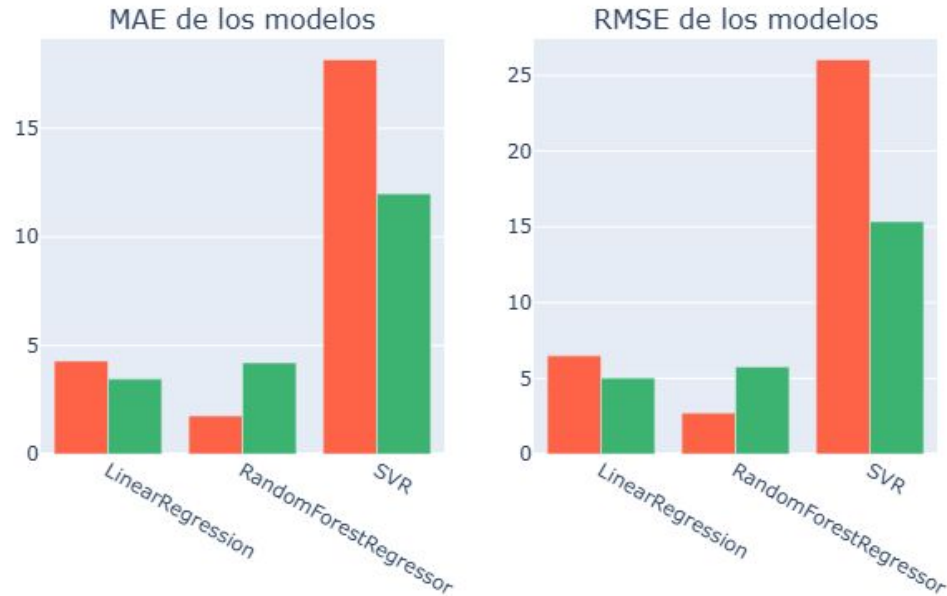
MAE = mean absolute error, RMSE = root mean square error

1. Regresión Lineal
 2. Random Forest Regressor
 3. Support Vector Regressor
- 

Modelos de predicción:

Regresión lineal (RL) - Random Forest Regressor (RFR) - Support Vector Regressor (SVR)

Resultados de los modelos por defecto



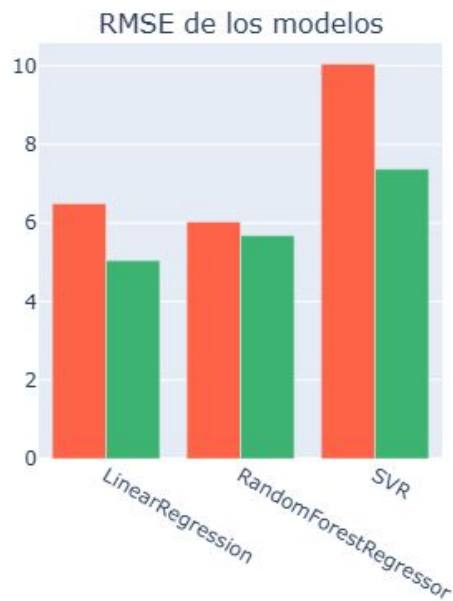
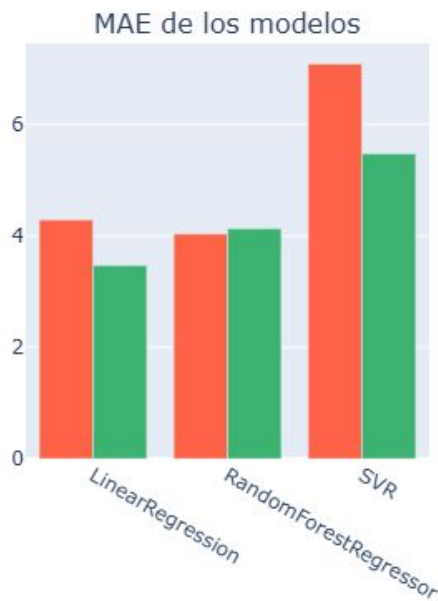
Rendimiento en testeo (según MAE y RMSE)

1. Regresión Lineal
2. Random Forest Regressor
3. Support Vector Regressor

Modelos de predicción:

Regresión lineal (RL) - Random Forest Regressor (RFR) - Support Vector Regressor (SVR)

Resultados de los modelos con búsqueda de hiperparámetros



Rendimiento en testeo (según MAE y RMSE)

1. Regresión Lineal
2. Random Forest Regressor
3. Support Vector Regressor


Modelos de predicción:

Regresión lineal (RL) - Random Forest Regressor (RFR) - Support Vector Regressor (SVR)

Predicciones de los modelos en el conjunto de test



Modelos ARIMA: ¿qué es?

- El modelo ARIMA (*AutoRegressive Integrated Moving Average*) es un método estadístico utilizado para modelar y predecir series temporales captando **dependencias lineales** y **patrones subyacentes** de los datos, como las *tendencias* o la *estacionalidad*.
 - Combinación de *tres componentes principales*:
 - *Auto-regresión* (p): relación entre una observación y sus valores pasados
 - *Integración* (d): Diferenciación de la serie temporal para que sea estacionaria
 - Promedio móvil (q): Relación entre el valor actual y los errores o residuos de valores pasados
- 

Modelos ARIMA: *modelos considerados*

- Consideraremos dos métodos principales: *ajuste manual* y *ajuste automático*.
- *Ajuste Manual*:
 1. (d) Test de Dickey-Fuller Aumentado para saber si la serie es estacionaria. Sino, diferenciar.
 2. (p, q) Truco de Auto-Correlación Parcial de residuos al cuadrado (el primer lag en el que se corta abruptamente)
 3. Comenzando desde $ARIMA(p, d, q)$, ajustar en base a las probabilidades por coeficiente y chequear que los residuos sean ruido blanco.
- *Ajuste Automático*:
 - Método *auto_arima* de la librería *pmdarima*

Modelos ARIMA: *ajuste manual*

- Consideramos $ARIMA(1, 1, 1)$ con residuos que se asemejan a ruido blanco.
- *Rolling prediction*:



Modelos ARIMA: *ajuste automático*

- Consideramos $ARIMA(4, 1, 4)$ con residuos que se asemejan a ruido blanco.
- *Rolling prediction:*



Redes Neuronales: MLP (Multilayer Perceptron)

- Los parámetros que mejor se ajustan al modelo son:
 - Activación: `identity` - Alpha: 0.001 - Hidden Layer Sizes: (200) - Solver: `lbfgs`

Predicciones del modelo MLP (gráfico completo)



Redes Neuronales: LSTM (Long Short Term Memory)

- Realizamos escalado con `MinMax` para la normalización de los datos en donde los valores estarán entre 0 y 1.
- Los valores deberán estar en secuencia con ventana de 30 lags.



Predicciones del modelo LSTM (gráfico completo)



Parte 3

Conclusiones sobre los modelos

- Conclusiones Finales

Resultado de los modelos

Gracias al trabajo realizado, pudimos ajustar los modelos de Regresión Lineal, Random Forest Regressor, SVR, ARIMA, MLP y LSTM a la serie temporal de Tesla para predecir el precio de cierre de la acción, siendo ARIMA y MLP los que mejor se ajustan y predicen el comportamiento de esta.

Modelo	RMSE
LR	5.042
RF	5.681
SVR	7.370
ARIMA	5.018
MLP	5.025
LSTM	11.303

