Uso del clasificador XGBoost para la predicción del mercado financiero Forex.

Mawrer Amed Ramírez Martínez

ms723737@iteso.mx

ITESO

Asesor

Dr. J. Guadalupe Olascuaga Cabrera

jolascua@iteso.mx

ITESO

Co-Asesor

Dr. Luis Fernando Gutiérrez Preciado

lgutierrez@iteso.mx

ITESO

*Resumen*—La investigación de distintas técnicas para interpretar y predecir movimientos en el mercado de divisas interbancarias Forex es muy extensa, va desde aplicación de métodos de control, pasando por identificación de patrones en gráficas, hasta la aplicación de redes neuronales y distintos tipos de algoritmos de aprendizaje automatizado. Múltiples investigaciones muestran resultados positivos en la aplicación de algoritmos de aprendizaje automatizado. Es un problema complejo con gran cantidad de variables, se han desarrollado distintitos enfoques para atacar el problema como el llamado análisis técnico, análisis fundamental y combinaciones de estos. En este trabajo se aborda el problema desde un análisis técnico, donde se estudian las características principales de los datos históricos de la tupla interbancaria EUR/USD recolectados desde el 1 de enero de 2014 hasta el 30 de mayo de 2020, el autor realiza un etiquetado de clasificación en base al porcentaje de cambio al cierre de cada periodo, se realizan experimentos de optimización para identificar los mejores valores de clasificación; como parte del análisis de los datos de entrada se utilizan herramientas estadísticas como la autocorrelación y correlación entre distintas características. Para la selección de nuevas características se utiliza la librería Technical Analysis (AT) Library para obtener indicadores a partir de la serie de tiempo original, en base a modelo de tipo Clasificador XGBoost se obtiene un subconjunto de características relevantes que serán utilizadas por los modelos finales. Posteriormente se entrena el modelo clasificador XGBoost para finalmente utilizarlo para predecir un subconjunto de datos de prueba, los cuales serán evaluados mediante simulación de trading (backstesting) obteniendo una media de resultados favorables.

*Palabras Clave*— ForReign Exchange Market, Exchange Rates, Machine Learning, Long Short Term Memory (LSTM), XGBoost, Prediction.

# INTRODUCCIÓN

El mercado de divisas es uno de los comercios de mayor impacto económico a nivel mundial. El comercio de negociación de monedas de todos los países (trading) logra generar flujos de hasta 5 billones de dólares diarios, el impacto de este flujo de capital está relacionado con aspectos macroeconómicos, sociales y políticos. En el mercado FOREX las divisas y su comportamiento son la base para la evaluación de gestión de riesgos y portafolios de inversión. El comportamiento de las divisas se clasifica dentro de los pilares más importantes a de la economía global; bajo esta afirmación surge la motivación de desarrollar una propuesta que proporcione respuesta a la siguiente pregunta: ¿Es posible anticipar, definir y obtener ganancias de los precios en las divisas en el mercado FOREX con el uso de algoritmos de aprendizaje automatizado?

Esta investigación hace uso de datos históricos del mercado interbancario FOREX y modelos de aprendizaje automatizando para la predicción de movimientos clasificados en tres tipos con el fin de identificar los mejores parámetros y manejo de datos.

El trading en los mercados Forex se entiende como la acción de compraventa de activos cotizados con mucha liquidez en el mercado (sobre todo, acciones, dividas y futuros) en un mercado financiero electrónico y regulado. Su objetivo es obtener un beneficio económico cuando la operación genera una plusvalía.

Con el fin de obtener resultados positivos y certeros en la predicción de movimientos del mercado, este articulo establece un modelo de clasificación basado en XGBoost. Para evitar la selección empírica y de forma aleatoria de los parámetros usados en los datos de entrada, se implementa una combinación de parámetros

El modelo es entrenado, validado y probado con datos históricos, la evaluación de las predicciones consiste en obtener las métricas de precisión, matrices de confusión y graficas del área bajo la curva, así como una ejecución de trading de forma simulada (backtesting) utilizando los valores predichos y evaluando el retorno de inversión.

# Trabajo Relacionado

Una de las razones por las que el trading se ha popularizado en los últimos años es debido a la entrada de nuevas tecnologías donde permiten operar en plataformas online especializadas y actúan como intermediarios en las operaciones o brokers.

Existen numerosas investigaciones relacionadas a la predicción de los mercados financieros. Algunas basadas en gráficas, las cuales intentan identificar patrones

(Hablar de análisis técnico, análisis fundamental, combinación, series de tiempo, características de series de tiempo, hablar de modelos y como hay evidencia que los mercados financieros están influenciados fuertemente por algoritmos).

# Trabajo propuesto

En esta siguiente investigación se propone el uso del modelo de clasificación XGBoost para la predicción de movimientos de cierre en el mercado interbancario Forex, para esta investigación se definieron 3 clases distintas a predecir [-1, 0, 1] donde la clase -1 será una clase para definir movimientos donde el precio de mercado al cierre fue menor con gran diferencia al precio de cierre de mercado al evento anterior, para la clase 0 se entiende que el movimiento sin gran cambio y la clase 1 corresponde a movimientos donde el precio de cierre fue mayor al precio de cierre del evento anterior.

Para el desarrollo de esta investigación se utilizó el siguiente ambiente: Python 3.7.3. Con las siguientes librerías de y versiones: Backtesting-0.1.4 [REFERENCIA], TA- 0.5.25 [REFERENCIA], xgboost-1.1.1 [REFERENCIA], keras- 2.3.1 [REFERENCIA], tensorflow- 1.14.0 [REFERENCIA].

Dentro de la investigación

(dejar en claro que el problema es complejo y que la cantidad de variables que afectan el tipo de cambio es muy grande, definir bien el tipo de problema), se …

Para el desarrollo de esta investigación se definió un flujo de trabajo que consta de distintas fases las cuales se muestran en la fig1.:



## Preparación de datos.

En esta etapa se obtienen y valida la integridad de los datos y los rangos en los que están distribuidos confirmando que no existan datos incongruentes en la serie de tiempo.

Los datos de entrada utilizados corresponden a dos conjuntos de datos principalmente: el primero con una frecuencia de cuatro horas en sus registros y el segundo de una frecuencia de un día entre registros. Ambos subconjuntos de datos fueron obtenidos de la plataforma Dukascopy (REFERENCIA) correspondientes al tipo de cambio EUR/USD, y van desde la fecha 1 de enero de 2014 hasta 30 de mayo de 2020.

En ambos casos para la creación del conjunto de datos inicial se utilizaron datos de compra y datos de venta correspondientes a las fechas descritas. Dando como resultado un conjunto de datos principal con un total de once columnas [*timestamp, Open\_ask, High\_ask, Low\_ask, Close\_ask, Open\_bid, High\_bid, Low\_bid, Close\_bid, Volume*].

### Características adicionales:

A partir de los datos de entrada de generaron indicadores estáticos de tipo estocásticos o de movimiento promedio.

Con el uso de la librería AT fue posible obtener 133 características adicionales.

Los datos utilizados de entrada utilizados son historiales del mercado Forex del banco Dukascopy [] del tipo de cambio EUR/USD (Euro – Dólar Americano).

### Características De Datos

El tipo de par interbancario utilizado es EUR/USD, fuente de datos. Recopilación y análisis de distintas frecuencias (1M, 1H, 4H, 1D). dando como resultado menor fluctuación y mayor ganancia en backstesting con el conjunto de datos de 1D.

Inputs del conjunto de datos principal: Time, Open, High, Low, Close y Volumen de tipo Compra y Venta.

Pct Change Diff, explicar autocorrelación.

Pct Change, explicar autocorrelación.

Hablar de los indicadores de tipo Osciladores, Moving Average, Estocásticos, Librería TA de indicadores técnicos.

## Extracción de características

En esta etapa se utiliza el conjunto de datos principal para ser la base de generación de nuevos datos de tipo movimiento promedio, estocásticos y estadísticos principalmente.

## Análisis y selección de características

Se utilizó el modelo de clasificación XGBoost como base para la selección de las características más importantes. XGBoost permite organizar las características relacionadas con su grado de influencia en el modelo de predicción. La importancia está basada en un puntaje denominado puntaje F (REFERENCIA) que indica la cantidad de veces que fue utilizada dicha característica para predecir un valor. Esta métrica muestra específicamente cuantas veces fue utilizada esta característica para el uso de árboles de decisión y promediada con el total de árboles de decisión existentes para este modelo.

Con el cual a partir de un análisis de threasure se determinó usar un total de X características.

## Entrenamiento del modelo

El conjunto de datos se dividió en tres subconjuntos correspondientes al proceso de entrenamiento, al proceso de validación y proceso de pruebas.

### Arquitectura de Modelos

Arquitecturas usadas en modelos de regresión, clasificación (LSTM, XGBoost, DNN)

### Algoritmos de Aprendizaje Automático

Como primera exploración se propuso predecir el precio de cierre con una

### Clasificación:

#### Long Short Term Memory (LSTM)

#### XGBoost (XGB)

#### DNN

Explicación de cada uno de ellos, separación de subsets de datos, oversampling, k-Fold, randomizacion, cross-validation.

## Validación del modelo

En la etapa de validación se utilizaron diferentes métricas como la exactitud absoluta, la matriz de confusión y el área bajo la curva para determinar los modelos que mejor predijeron los subconjuntos de datos correspondientes a validación y pruebas.

## Evaluación de predicciones

En esta etapa se utilizó la librería Backtesting de Python para la simulación de trading en base a las predicciones realizadas a partir del modelo XGBoost.

### Evaluación De Modelos

Una vez el modelo XGBoost es entrenado es necesario evaluar

### Métricas de medición para evaluar el performance de los modelos con datos de pruebas.

### Uso de la librería backtesting para evaluación de resultados predichos.

# Experimentos y Resultados

## Experimentos de regresión (Close), (PCT Change) (PCT Change Diff).

## Experimentos de Clasificación

## Optimización de Backtesting.

En los experimentos realizados para backtesting se

## Experimentos de Acc y profundidad utilizada.

## Identificación de características relevantes con modelos XGBoost en distintos conjuntos de datos (1H, 1D) y distintas profundidades.

## 

## Resultados

Resultados obtenidos a partir de la evaluación de modelos con backtesting. Algo breve.

# Conclusiones

La mejor dataset para predicción fue de 1D y 4H.

Mediante el uso del clasificador XGBoost fue posible

Referencias

1. G. O. Young, “Synthetic structure of industrial plastics (Book style with paper title and editor),” in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.