Desafío final

Predicción de acciones en base a sentiment analysis

Objetivo:

- Predecir el retorno excedente de una acción ( NFLX) respecto al indice NASDAQ 100 en base a recopilación de información de internet y sentiment analysis

- Inspiración: Sentiment analysis of Twitter data for predicting stock market movements (V. S. Pagolu, K. N. R. Challa, G. Panda, and B. Majhi)

Recolección de datos

2 fuentes:

1. Scraping de Twitter mediante Twint.

2. Recopilación de datos de Reddit mediante su API.

Un año de datos, búsquedas de “NFLX” o “Netflix”

Limpieza y preprocesamiento

Extraemos URLs, hastags, emoticones del texto, y lo preparamos para el sentiment analysis. Intentamos distintos métodos:

1. Expresiones regulares
2. Tokenizers
3. Filtros manuales (stopwords, longitud de palabras)
4. Lemmatizer, stemmer

Sentiment analysis

Usamos:

1. TextBlob
2. Vader Sentiment Intensity Analyzer

Devuelve tres scores (positivo, negativo, neutro).

El dataset final tiene una entrada por día, con tres features correspondientes al promedio de cada score para los textos de cada día. La variable target era si el excess return del día siguiente era positivo o negativo

Clasificación

Probamos múltiples algoritmos (Random Forest, XGBoost, Regresión Logística, Support Vector Machine) e hicimos una búsqueda de hiperparámetros en una RandomizedSearch. Incluimos también dos hiperparámetros ‘manuales’: si considerábamos o ignorábamos valores neutros, y el tamaño de la ventana de días

Usamos split en series de tiempo para cross-validation.

El mejor resultado obtenido fue un Random Forest, con ventana de 3 días, usando TweetTokenizer y stopwords. El ROC AUC score fue de 0.587 en train y 0.792

Mejoras

- Ampliar rango de tiempo de datos para cubrir más ciclos de mercado

- Desarrollador clasificadores propios. Entrenar clasificadores independientes para distintas fuentes de datos

- Ampliar features