Universitat Oberta de Catalunya

uoc.edu

Estimación del Coste del Gas en transacciones de Ethereum mediante Deep Learning

Antonio Arias Sánchez
Máster Univ. en Ing. de Telecomunicación
TFM - Área de Telemática

Dr. José López Vicario
Dr. Xavier Vilajosana Guillén

Enero 2022

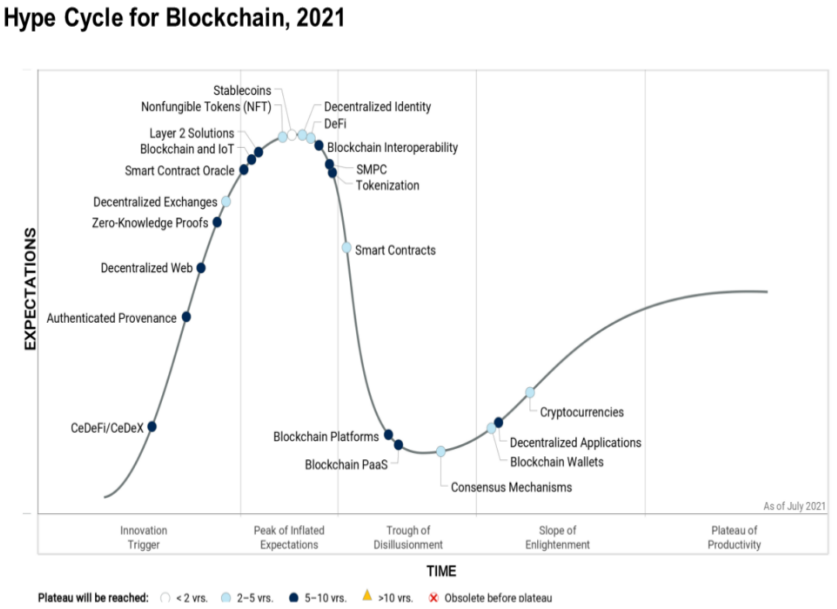
Índice

Objetivos

- Iniciarse en **IA, Machine Learning, Deep Learning**.
  Manejo de principales frameworks: TensorFlow y Keras
- Iniciarse en **Blockchain**, Bitcoin y **Ethereum**
- Identificar y Estudiar un problema de Blockchain y abordarlo mediante Deep Learning: **Estimación del Precio del Gas en Ethereum**
- Revisión de otros **trabajos existentes** en problemas similares
- Obtención y Preparación de un **Dataset** adecuado
- Definición, entrenamiento y test de **Red Neuronal.** Análisis y Comparativa de Resultados.
- Constituir **TFM** para el MU Ing. de Telecomunicación

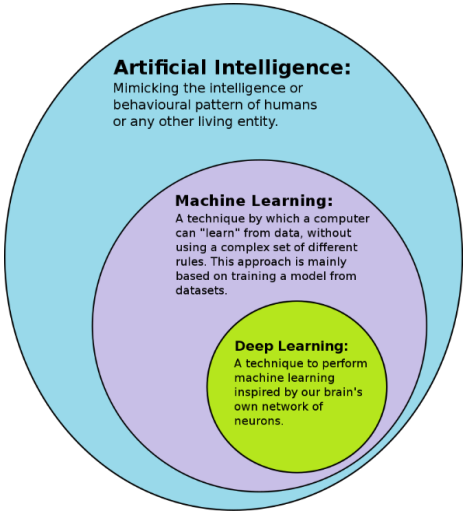
Contexto

- **Blockchain** comienza a ser **tecnología establecida** y utilizada.
  Bitcoin y Ethereum, líderes de uso.
- Al aumentar el uso, problemas de **escalabilidad**: lentitud y alto coste transacciones
- Comprender el **coste del Gas**, problema de interés práctico en Ethereum
- Problema interesante para **abordar con Deep Learning**
- Existen algunos trabajos, pero mayoría orientados al precio del Ether vs USD, no del Gas.



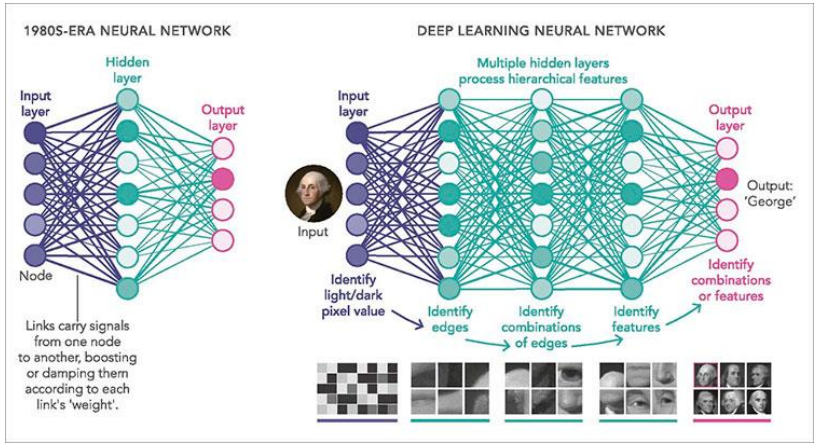
IA y Machine Learning

- **IA** es una **disciplina veterana**, con períodos sucesivos de éxitos y euforia, seguidos de estancamientos y pérdidas de interés.
- En la actualidad, **nuevo auge** gracias a los espectaculares resultados obtenidos mediante Machine Learning y **Deep Learning**.
- Nace en la década de los **1950's**, primeros esfuerzos en la **IA simbólica**, sistemas acotados, heurística y Sistemas Expertos.
- Desde los años **1990's**, mejores resultados proceden de la **IA conexionista** y el Machine Learning.
- **Machine Learning** persigue que un sistema sea capaz de **resolver problemas analizando** largas colecciones de **ejemplos** de donde extraer patrones.
- Aproximaciones:
  - Aprendizaje **supervisado**. Ejemplos de entradas y salidas deseadas
  - Aprendizaje **no supervisado**. Sólo entradas de donde extraer patrones
  - Aprendizaje **reforzado**. Entradas dinámicas y señal de feedback o refuerzo.



Deep Learning

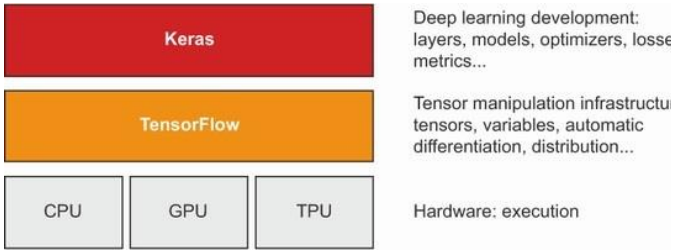
- **Deep Learning**, técnicas de Machine Learning que emplean **Redes Neuronales Artificiales** multi capas. La **profundidad** es variable, se refiere al número de **capas ocultas** empleadas. Mayor que uno para conseguir resultados significativos.
- **Áreas de aplicación** características:
  - Reconocimiento de Voz (ASR)
  - Procesado de Lenguaje Natural (NLP)
  - Visión por Computador (CV)
  - Clasificación y Predicción de series
- **Tipos de Redes Neuronales** más usados:
  - Perceptrón Multicapa. Densamente interconectado.
  - Redes Convolucionales (CNN). Imágenes y Visión.
  - Redes Recurrentes (RNN). Series temporales



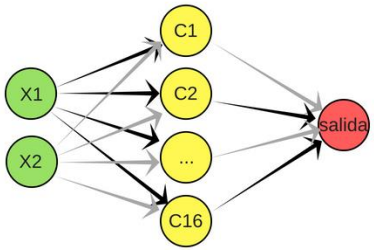
- Causas **avance** actual de Deep Learning**:**
  - **Potencia de Cómputo**. GPU, TPU, Cloud.
  - **Big Data**. Disponibilidad grandes datasets.
  - **Algoritmos** y Frameworks abiertos.

Tensor Flow y Keras

- **Tensor Flow**, librería de código abierto desarrollada por Google desde 2017, para construir todo tipo de sistemas de Machine Learning. Basada en flujo de trabajo sobre Tensores. **Python**, lenguaje de programación principal.
- **Keras.** Interfaz sobre el anterior, usado para construir modelos de Redes Neuronales.
- **Google Colab.** Entorno de desarrollo y ejecución en nube de Google para proyectos de Machine Learning. Permite ejecutar cuadernos Jupyter en Python y emplear GPUs y TPUs.



- **Keras** permite una sencilla implementación de Redes Neuronales:



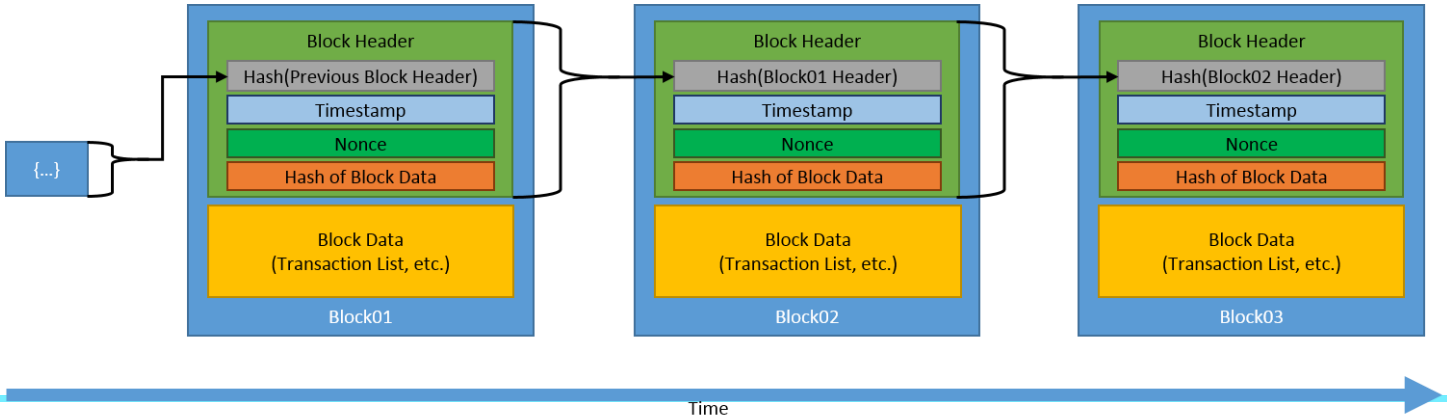

```
model = Sequential()
model.add(Dense(16, input_dim=2, activation='relu'))
model.add(Dense(1, activation='sigmoid'))

model.compile(loss='mean_squared_error',
              optimizer='adam',
              metrics=['binary_accuracy'])

model.fit(training_data, target_data, epochs=1000)
```

Blockchain, Bitcoin, Ethereum

- **Blockchain** es una tecnología que permite crear un **registro de datos público**, inmutable, que aumenta con el tiempo, y cuya validez puede verificarse **sin** necesidad de una **autoridad central**.
- Nace en 2009 con **Bitcoin**, donde se utiliza para implementar una divisa electrónica.
- Sus casos de uso se amplían con la aparición de las **Aplicaciones Distribuidas**, introducidas en **Ethereum** en 2013. Éstas se ejecutan sobre una Máquina Virtual y su estado se mantiene dentro de la Blockchain.



Bloques y Transacciones en Ethereum

- **Cuentas,** direcciones desde las que se interacciona con la Blockchain:
  - **Controladas externamente**. De usuarios, controladas por claves privadas.
  - **De contratos**. Controladas por código. Se activan al recibir una transacción.
- **Transacciones**:
  - Unidad de interacción con la red. Iniciada por usuario desde una cuenta. Puede **intercambiar divisa** (Ether), y/o **invocar un contrato** (paso de parámetros).
  - Tienen **coste**, para recompensar mineros, y evitar bucles infinitos.
  - Las iniciadas por un contrato, se denominan **mensajes**.
  - Son **confirmadas** sólo tras quedar incorporadas en un bloque.
- **Bloques**:
  - Son generados por los mineros, agrupando transacciones válidas pendientes.
  - Siguen un orden secuencial, desde el bloque inicial o génesis.
  - Se generan a un ritmo medio de 250-300 bloques / hora.

Coste de Transacciones. Precio del Gas.

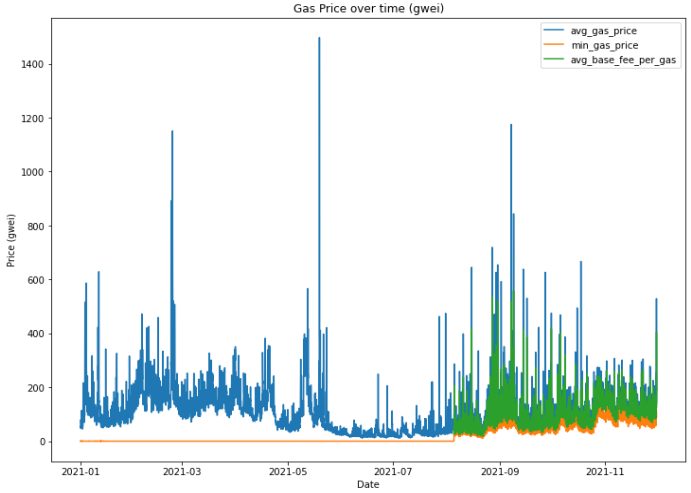
- Las transacciones consumen **Gas**, que a su vez **se compra mediante Ether**
- Al iniciar una **Transacción**, un **usuario debe indicar**:
  - **Límite de Gas**. Máximo que debe emplearse. Si se supera, la transacción se aborta.
  - **Precio del Gas**. Precio en Ether por unidad de Gas que se desea pagar.
- Al formar un nuevo bloque, los mineros toman las transacciones con mayor precio del gas primero. La **elección óptima** debe ser ofrecer el precio más bajo posible que garantice la inserción.
- **Desde la EIP-1559** (Ago 2021), la red calcula un **Precio Base del Gas**, y el usuario puede indicar, en lugar del Precio del Gas, la prima máxima sobre el precio base, y el máximo absoluto a pagar.
- El **elevado coste de las transacciones** continúa siendo un problema. Algunas tendencias:
  - Surgimiento **Blockchains 3ª generación** (Cardano, Polkadot)
  - **Ethereum 2.0**. Futuras Shard Blockchains
  - Paso de Proof-of-Work a Proof-of-Stake

Otros trabajos sobre Deep Learning y Blockchain

- Existen **numerosos trabajos** que tratan de **predecir el precio del Ethereum** -u otras crypto divisas- **frente al USD**, de manera similar al de otros activos financieros.
- Algunos **trabajos** aplicados específicamente **a la red** Ethereum, proponen lo siguiente:
  - Predicción de éxito o fracaso de transacciones. Clasificación binaria.
  - Tiempo medio en bloques de confirmación de transacción. Modelo de clasificación en varias clases.
  - Predicción del precio mínimo del Gas, empleando un modelo de regresión.
  - Predicción del precio mínimo del Gas, empleando RNN.
- **El presente trabajo**:
  - Predicción del precio mínimo del Gas, empleando varios modelos ML y DL.
  - Método de Ventana deslizante.
  - Empleo de amplio conjunto de features.

Obtención de Dataset

- Se encuentra en **Google Big Query Ethereum** una buena fuente, fácilmente accesible, con todo el histórico de la Blackchain Ethereum.
- Se crean **scripts** para capturar el histórico del Precio del Gas, y 12 features relacionadas, por unidades de tiempo.

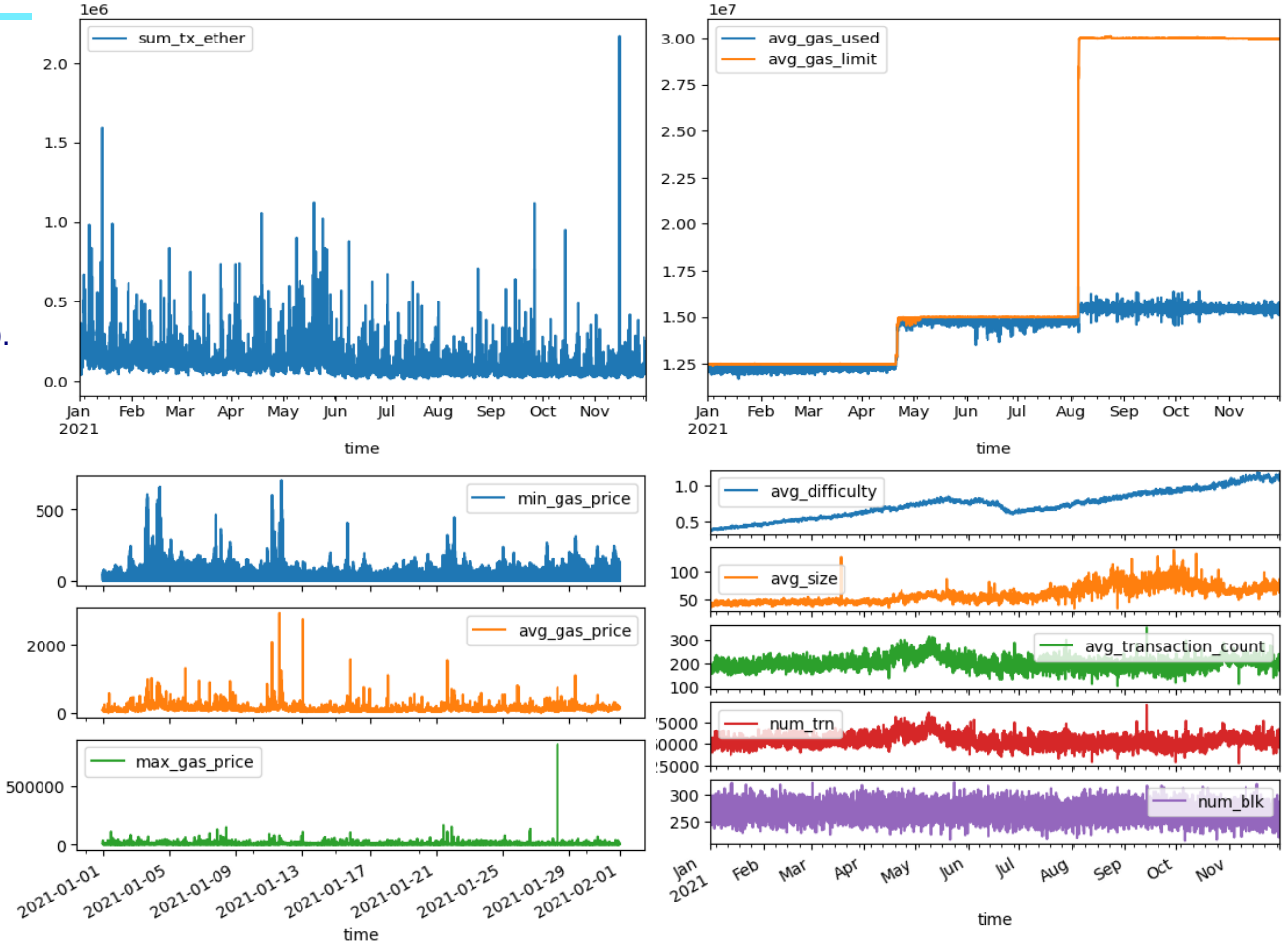


Dataset. Features.

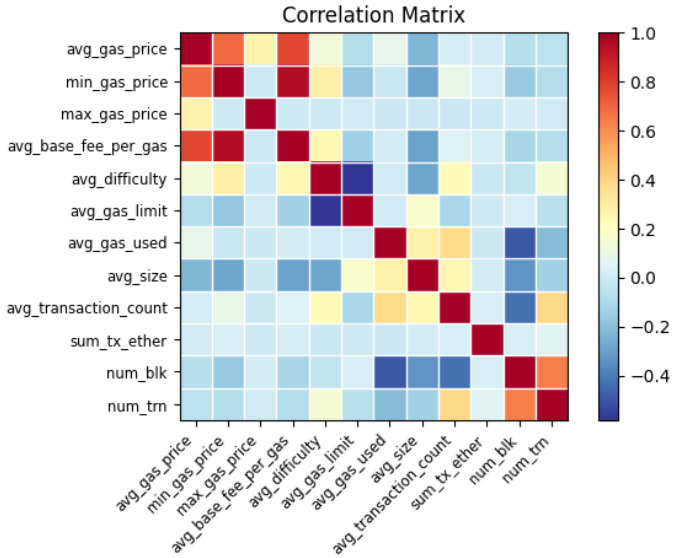
Además del precio, otras **features** capturadas:

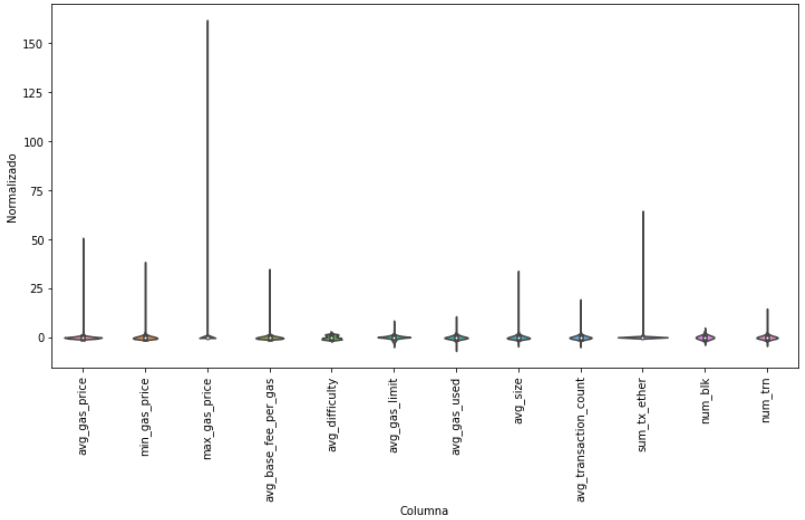
- Dificultad de la red.
- Tamaño del Bloque en kB
- Núm. de transacciones / tiempo.
- Núm. de bloques / tiempo
- Uso de gas por bloque.
- Límite de gas por bloque.
- Ether total transferido.

Todas ellas reflejan la ocupación de la red en cada instante.



Análisis Preliminar y Limpieza de Datos

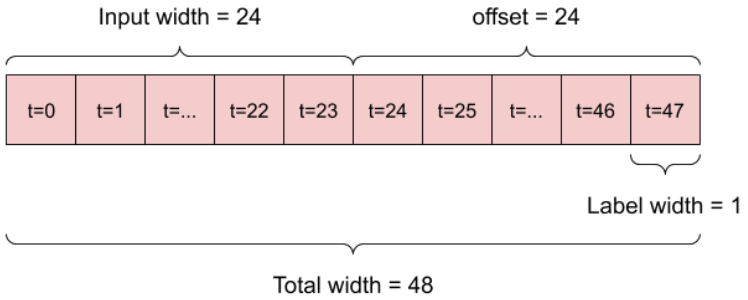


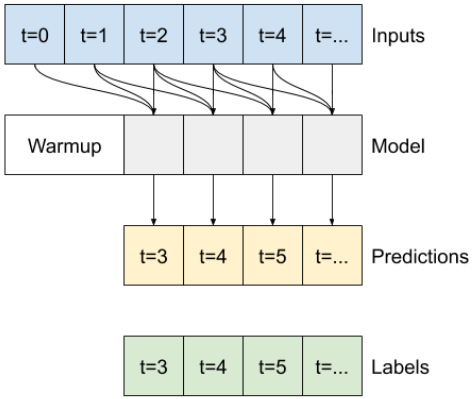


- Baja correlación en general entre features en ventanas de 1 minuto, que aumenta para 1 hora.
- Mucha dispersión (outliers) en algunas features, precio máximo y medio, y ether transferido.
- Se eliminan outliers y datos faltantes en los datasets obtenidos

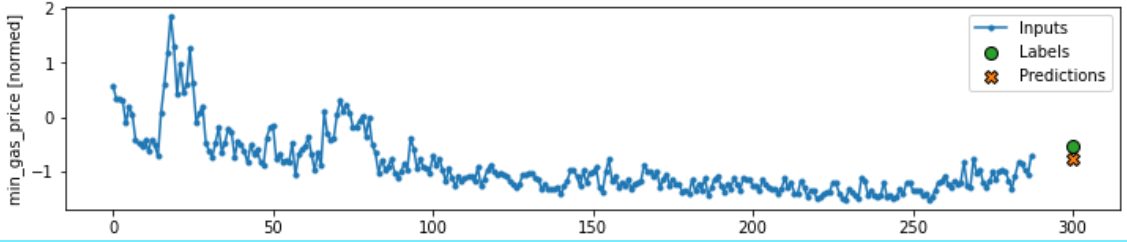
Método de la Ventana Deslizante

- Habitual en predicción de series temporales.
- Parámetros:
  - Longitud entrada
  - Offset
  - Longitud predicciones





- Se toma una secuencia de todas las features para generar la predicción como valor puntual –o secuencia- de los valores futuros distanciados el offset.
- En el trabajo, se han empleado ventanas de:
  - 24 horas, muestras cada 5 minutos, y offset de 1 hora
  - 12 horas, muestras cada minuto, y offset de 1 hora



Creación y Entrenamiento de Modelos

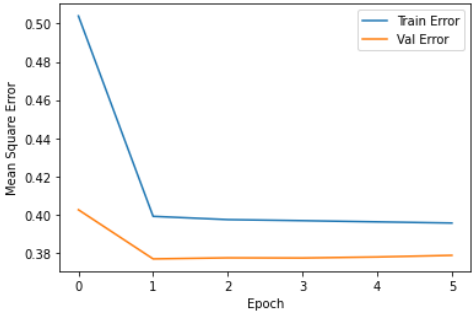
- Se han **creado y entrenado 3 modelos** diferentes, además de 1 baseline:
  - **Baseline**. Predice el último valor de la muestra de precios.
  - Modelo de **Regresión Lineal**. Modelo básico de Machine Learning, que se puede generar en Keras mediante una capa sin función de activación.
  - **Perceptrón Multicapa**. Red Neuronal Densa de dos capas, y función de activación ReLu.
  - **Red Neuronal Recurrente**. Debería ser la más adecuada para una serie temporal, ya que mantiene internamente estado de lo ocurrido anteriormente. Basada en capa LSTM.

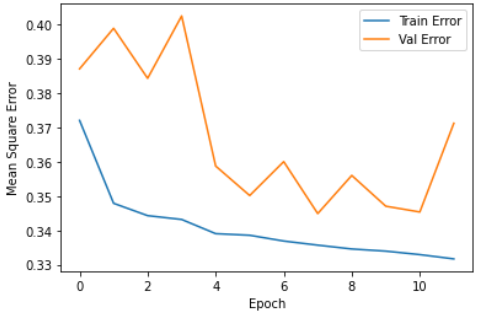
- Para el **entrenamiento y pruebas**, se ha se realizado la división de los datos en los habituales tres datasets, de entrenamiento, validación y test.

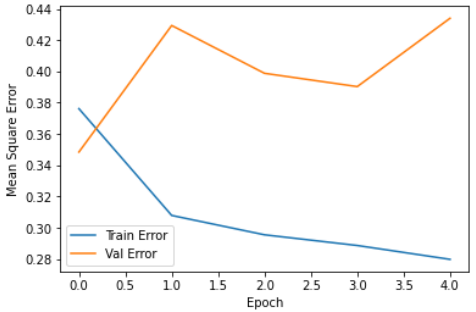
- Cada dataset está compuesto de **batches** obtenidos aplicando sucesivamente la ventana deslizante a la serie temporal de partida.

Resultados de Entrenamientos

Empleando ventanas de 24 horas, muestras cada 5 minutos, y offset de 1 hora.

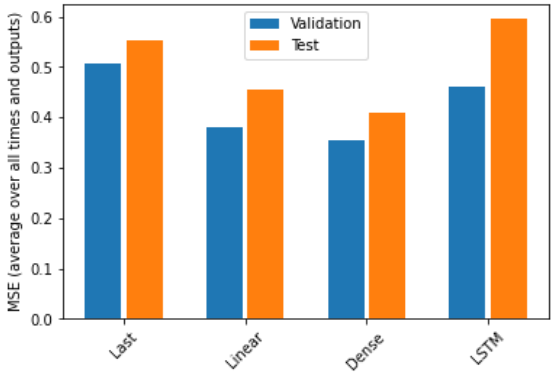


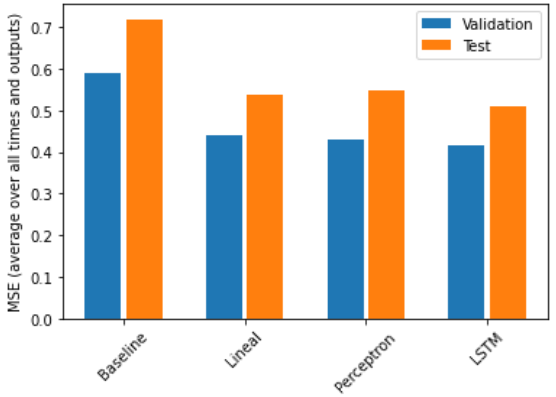




De izquierda a derecha, resultados para Modelo de Regresión Lineal, Perceptrón y RNN.
Se puede apreciar como el modelo RNN en este caso presenta peores resultados que el perceptrón, pues en seguida comienza a experimentar overfitting.

Comparativa de Resultados





De arriba abajo, vemos la comparativa para los casos de:

- Ventana 288 muestras (24 horas, muestra cada 5 minutos) y offset de 1 hora
- Ventana 720 muestras (12 horas, muestra cada minuto) y offset de 1 hora

En el último caso, apreciamos como los modelos más complejos ofrecen sucesivamente mejores resultados.

En el primer caso, comentado anteriormente, el modelo RNN se comporta bastante peor. Sólo al aumentar el número de muestras y la resolución temporal se consigue mejora con el mismo.

Conclusiones

- Se ha planteado y **ejecutado** con éxito **el ciclo de un Proyecto de Deep Learning** típico para el problema del Precio del Gas.
- Se comprueba que las **Redes Neuronales** son aplicables al problema, y **consiguen mejor resultado** que modelos más simples, aunque con capacidad limitada, pues se aprecia enseguida bastante overfitting.
- Esto sugiere poca capacidad predicativa en las features analizadas. **Se proponen otras features** que podría ser interesante analizar, como el Pool de transacciones pendientes o el precio del Ether u otros contratos.
- Se ha constatado la **importancia de la escala de tiempo** en el estudio**. Algunas features** presentan correlación a largo plazo**; a corto plazo**, en cambio, **tienen capacidad predictiva**, pues son indicadores de inestabilidades temporales de la red.
- Se comprueba que en un proyecto de Data Science, la **obtención y procesado de Datos** es la parte que más dedicación requiere.
- Igualmente, se constata la **importancia de adquirir familiaridad con el Dataset para elegir** el tipo de Red Neuronal sus **meta parámetros**.
- Finalmente, se señalan **algunas posibles vías de ampliación** del estudio, como la relación entre el precio del Gas y el retardo, o analizar el algoritmo de cálculo del precio base incorporado en la EIP-1559 y compararlo con los resultados obtenidos mediante Deep Learning.

Estimación del Coste del Gas en transacciones de Ethereum mediante Deep Learning

¡Muchas Gracias!

Antonio Arias Sánchez
Máster Univ. en Ing. de Telecomunicación
TFM - Área de Telemática

Dr. José López Vicario
Dr. Xavier Vilajosana Guillén

Enero 2022