# Bitcoin Price Prediction Pipeline

**Berkan Yapıcı  Yağmur Eren**

## Problem Description

The project seeks to address the challenging task of predicting daily Bitcoin prices through the development of an advanced machine learning system. In pursuit of this objective, the primary data source will be the Binance API, a comprehensive platform that supplies historical daily Bitcoin price data. The central challenge at hand is the forecasting of future Bitcoin prices, a problem that inherently involves unraveling intricate patterns embedded within the historical data and incorporating pertinent features.

Bitcoin, being a highly dynamic and volatile cryptocurrency, poses a complex prediction problem. The goal is to create a robust model capable of capturing the underlying trends, patterns, and dependencies in the historical price movements. By leveraging machine learning techniques, specifically Multi-Layer Perceptron and (MLP) and Long Short-Term Memory (LSTM) networks, the project aims to provide a reliable tool for forecasting the future trajectory of Bitcoin prices.

## Tools

The project is designed to harness a versatile set of tools, strategically chosen to ensure efficiency and effectiveness in building the machine learning (ML) system. Each tool serves a specific purpose, collectively contributing to the seamless development and deployment of the Bitcoin price prediction pipeline. AWS lambda functions are used instead of github actions for daily feature pipeline. Binance API is restricts the API calls coming from USA. The location of lambda functions can be adjusted. Another issue was lambda functions has a size restriction of 50 mb. Docker images are used to create containers for function. This approach overcomes the size requirements. Function special docker image is uploaded to AWS ECR elastic container registry and AWS EventBridge is used for activation of lambda functions.

1. **TensorFlow:**
    - **Role:** Core Machine Learning Library
    - **Description:** TensorFlow, an open-source ML library, will be at the heart of the project, driving the implementation of the Long Short-Term Memory (LSTM) model. Renowned for its flexibility and scalability, TensorFlow provides a robust foundation for training and deploying complex neural network architectures.
2. **Hopsworks:**
    - **Role:** Feature Store and Model Registry
    - **Description:** Hopsworks stands as a comprehensive platform, seamlessly integrating the feature store and model registry into the ML pipeline. This tool

streamlines the process of managing features and registering models, ensuring a unified and organized approach to feature engineering and model development.

3. **AWS Serverless Lambda Functions:**
    - **Role:** Execution Environment for Code
    - **Description:** Serverless Lambda functions are employed to execute code without the need for dedicated servers. This serverless paradigm enhances flexibility, scalability, and resource efficiency, aligning with modern cloud-native architectures.
    -
4. **Docker:**
    - **Role:** Containerization for Lambda Functions
    - **Description:** Docker containers are utilized to package and deploy Lambda functions seamlessly. This containerization ensures consistency across different environments, simplifying the deployment process and enhancing the portability of the solution.
5. **EventBridge:**
    - **Role:** Event-Driven Function Invocation
    - **Description:** AWS EventBridge facilitates event-driven architecture by enabling seamless communication between various components. It plays a crucial role in orchestrating the invocation of Lambda functions based on specific events or triggers.
6. **Google Colab**
    - **Role**: Machine Learning Model Training
    - **Description**: Google Colab is employed as the primary platform for training machine learning models, specifically the LSTM model for Bitcoin price prediction. Leveraging the cloud-based infrastructure provided by Google Colab, the project taps into the computational power of T4 GPUs to expedite the training process. Google Colab's collaborative and interactive environment facilitates seamless experimentation with different model architectures and hyperparameters. By utilizing Google Colab, the project ensures efficient model training, benefiting from a cloud-based, GPU-accelerated environment that enhances the speed and effectiveness of the LSTM model development.
7. **Hugging Face Spaces:**
    - **Role:** Interface Development
    - **Description:** Hugging Face Spaces plays a central role in fostering collaborative and dynamic development of the user interface (UI) for the Bitcoin price prediction system. As a collaborative platform, Hugging Face Spaces enables team members to collectively contribute to the UI development process. The platform's interactive and collaborative nature facilitates real-time collaboration, allowing developers, designers to work seamlessly on the UI.

## Data

The primary data source is the Binance API, which provides historical Bitcoin price data. The dataset will include features such as open price, high price, low price, close price, volume, and other relevant metrics. Data collection will involve periodic retrieval of daily price information through the Binance API.

## Structure of the Project

The project's structure, Figure 1, is composed of four primary components: a feature pipeline, a training pipeline, a daily feature pipeline, and a user interface. The feature pipeline is responsible for data acquisition, utilizing the Binance API to retrieve data. Subsequent preprocessing tasks are performed on this data, ensuring that the features are suitably arranged before being stored in the feature store.

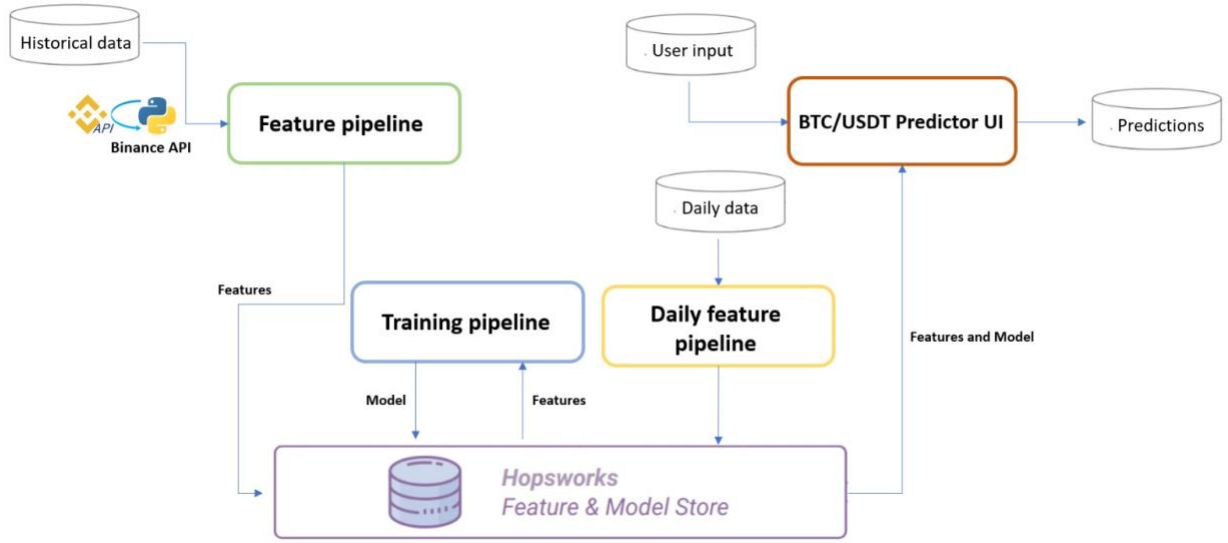


*Figure 1 A diagram of the final pipeline*

In the training pipeline, the LSTM model is trained using the preprocessed features from the feature store. The model is registered for future use.

The daily feature pipeline functions to update the feature store with the most recent data. It takes the current day's price, applies the same preprocessing as done in the feature pipeline, and then stores this updated information in the feature store.

Lastly, the user interface plays a critical role. It captures user input regarding the desired prediction period, specifying how many days ahead the prediction should extend. Utilizing the up-to-date data from the feature store and the registered LSTM model, the system then generates and displays a graph representing the predicted prices for the specified period.

## Methodology and Algorithm

In the pursuit of accurate and effective time series prediction for Bitcoin prices, a variety of models were explored, with notable success found in MLP (Multi-Layer Perceptron) and LSTM

(Long Short-Term Memory) architectures. The proposed methodology centers around the application of LSTM, a specialized type of recurrent neural network (RNN), renowned for its proficiency in capturing long-term dependencies within sequential data. Given the inherent sequential nature of time series data, LSTM stands out as an apt choice for the prediction task.

The LSTM model will be employed to analyze historical Bitcoin price data, leveraging its ability to discern intricate patterns and dependencies over extended temporal periods. Feature engineering will play a crucial role in enhancing the model's understanding of the underlying dynamics, incorporating relevant metrics such as open price, high price, low price, close price, and volume. This holistic approach ensures that the LSTM model is equipped to grasp both short-term fluctuations and long-term trends within the Bitcoin market.

The hyperparameter tuning process is a pivotal step in optimizing the LSTM model's performance. A small subset of the dataset will be dedicated to this task, facilitating the fine-tuning of parameters such as the number of layers and nodes per layer. This meticulous tuning aims to strike a balance between model complexity and efficiency, tailoring the LSTM architecture to the specific characteristics of the Bitcoin price time series.

The training phase will be conducted on Google Colab, utilizing T4 GPUs to expedite the process. This infrastructure choice optimizes training time, allowing for efficient exploration of various architectural configurations. The training duration is expected to be less than 12 hours, ensuring a timely and resource-efficient development cycle.

By combining the strengths of LSTM architecture, thoughtful feature engineering, and strategic hyperparameter tuning, the proposed methodology aims to deliver a robust and accurate time series prediction model for Bitcoin prices. The chosen approach acknowledges the unique challenges posed by cryptocurrency markets and seeks to harness the power of deep learning to navigate the intricacies of Bitcoin price dynamics.

## Results

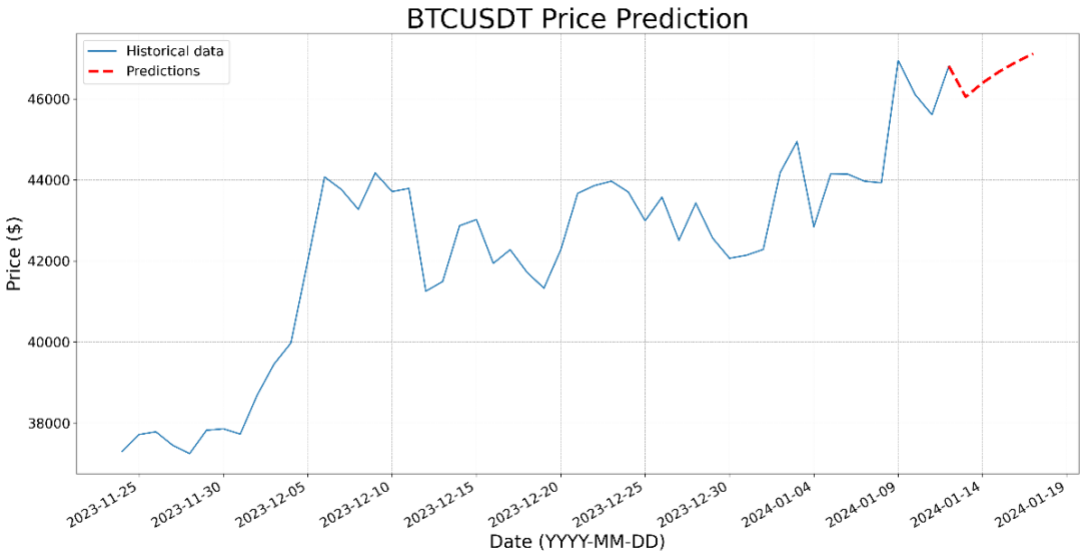


*Figure 2 Example graphic from the results*

The project encompasses the development of a Bitcoin price prediction tool, employing LSTM networks for accuracy. This tool is designed to allow users to specify the number of days they wish to predict into the future via a user interface. When a user inputs a desired timeframe (in days), the tool generates a detailed graph given in Figure 2. This graph integrates both the historical price data from the past 50 days and the predicted prices for the forthcoming days. In terms of performance, the model demonstrates commendable efficacy in short-term predictions.