DSCI 6007 Final Project Report

Forecasting Inflation with an LSTM Neural Network Through an AWS Data Engineering Pipeline

Team 2

Jafar Vohra

April 23, 2024

Abstract:

In today's data-driven landscape, the efficiency of data engineering pipelines is paramount for organizations across industries. This report details the design and implementation of a robust data engineering pipeline, aimed at facilitating data ingestion, storage, processing, and machine-learning model deployment for economic forecasting. Leveraging AWS Sagemaker, the project commenced with data storage and ingestion tasks, utilizing Python and Pandas for seamless handling of data. Through meticulous data cleaning and exploration, significant economic trends and relationships were uncovered, paving the way for model development.

The predictive analysis phase involved transforming the data into sequences for LSTM-based time-series forecasting. Model training and evaluation demonstrated the efficacy of the approach, with the model accurately forecasting Core CPI trends. The deployment phase ensured accessibility by deploying the model as an endpoint on AWS Sagemaker, enabling real-time predictions of Core CPI and its year-over-year percentage change.

The project's culmination underscores the pivotal role of robust data engineering pipelines in deriving actionable insights and driving informed decision-making. Looking ahead, the report highlights opportunities for further refinement and scalability, emphasizing the continuous evolution of data-driven methodologies. Through the integration of advanced technologies and scalable architectures, the pipeline lays a foundation for future endeavors in economic forecasting and beyond.

Introduction:

In the data-driven world, organizations across various industries rely on efficient data engineering pipelines to streamline data ingestion, storage, processing, and consumption. Such pipelines are the backbone for deriving actionable insights, driving decision-making, and delivering value to stakeholders. The objective of this project is to design and implement a robust data engineering pipeline that facilitates these essential functions, culminating in the deployment of a machine-learning model for real-world application.

Methodology and Results:

The leading task was to store the data safely and accessibly. The simplest way to accomplish this was by uploading the data to an AWS S3 bucket. After gathering data for the project and saving it to a CSV file, it could be uploaded to a preformed bucket. The code for this task was written in Python while utilizing the sagemaker module on the AWS Sagemaker Notebook instance. The code simply created a sagemaker session before defining the bucket name and path within the bucket in which the data will be saved. It then used the "upload_data" method with the previously defined variables to complete the upload.

Now with the data storage complete, the task of data ingestion could be tackled. This process mirrored the data storage task, however, it could be modified to handle batch data processes in the future if the project is scaled further. The CSV data file was read with the help of the Pandas library after importing it and the rest of the necessary dependencies for the notebook. The data was then ready to be processed and formatted for the upcoming modeling portion of the project.

After cleaning the data of missing values, there were 334 records with 12 features for each data record. The types of data were observed, which signals that a change in the datatype for the “date” feature was necessary. Once handled, a copy of the processed Pandas DataFrame was made for the upcoming exploratory data analysis.

A series of transformations were performed to prepare the data for analysis. To better understand the trends, the percentage change in the Core CPI month-over-month (MoM) and year-over-year (YoY) was calculated. Time-based features such as year, quarter, and month from the date feature were extracted to complete this step, allowing us to analyze trends and patterns over specific periods. To visualize the data, we created line plots to examine trends of different indicators, including the unemployment rate, personal saving rate, M2 (a measure of money supply), disposable income, and the fed rate. The plots provided insights into how these factors changed over nearly three decades. Also, box plots were used to explore the Core CPI year-over-year percentage change by month and quarter. This helped identify any significant variations or outliers across different periods. Bar plots showing the standard deviation for the Core CPI by month and quarter further depicted the variability of this indicator.

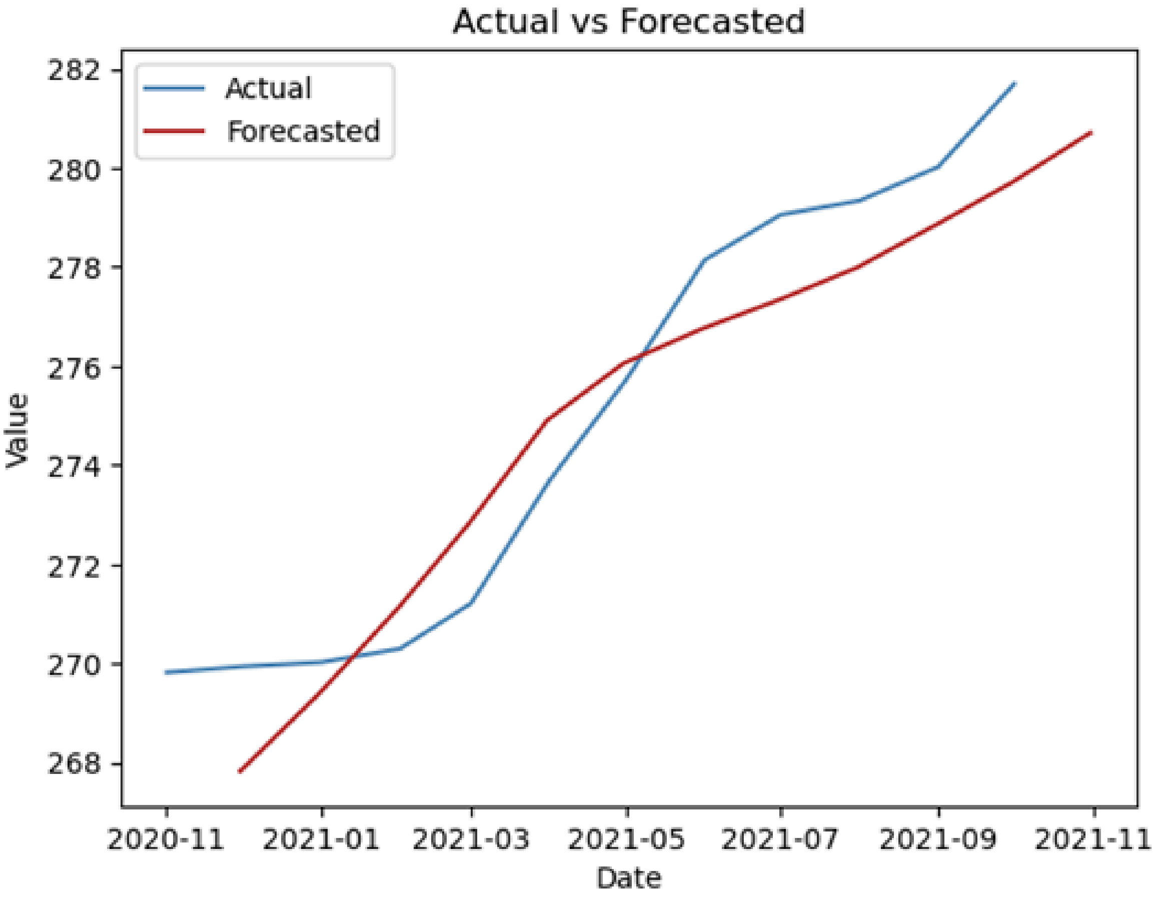
For accurate time-series analysis, the data must be stationary, meaning statistical properties don't change over time. The Augmented Dickey-Fuller (ADF) test was leveraged to check for stationarity in each indicator. If a series was not stationary, differencing, a method that calculates the difference between consecutive observations, was applied to stabilize it. This transformation was repeated as needed to ensure the data met the requirements for time-series analysis.

Once the data was stationary, Granger's Causality Test was utilized to explore potential causal relationships between the indicators and the Core CPI. This test helps determine if one feature is statistically significant in predicting another, providing valuable insights into the interplay between different economic factors. Through this test, it was determined that the real effective exchange rate, ten-year treasury yield, and fed rate were not significant indicators, thus they were removed from the data that would eventually train the model.

Once the most important features were selected, the data was scaled through min-max normalization. This step led to the data being transformed into sequences, following a sliding-window approach, which allowed the creation of input-output pairs to be used in the LSTM-based time-series forecasting model. This process provided the foundation for the predictive analysis of Core CPI trends. The 1994-2019 period was used to train the model while 2019-2020 was used to evaluate model performance. After being reshaped and stacked, the sequences of data were split. This completed the data processing for the model training data.

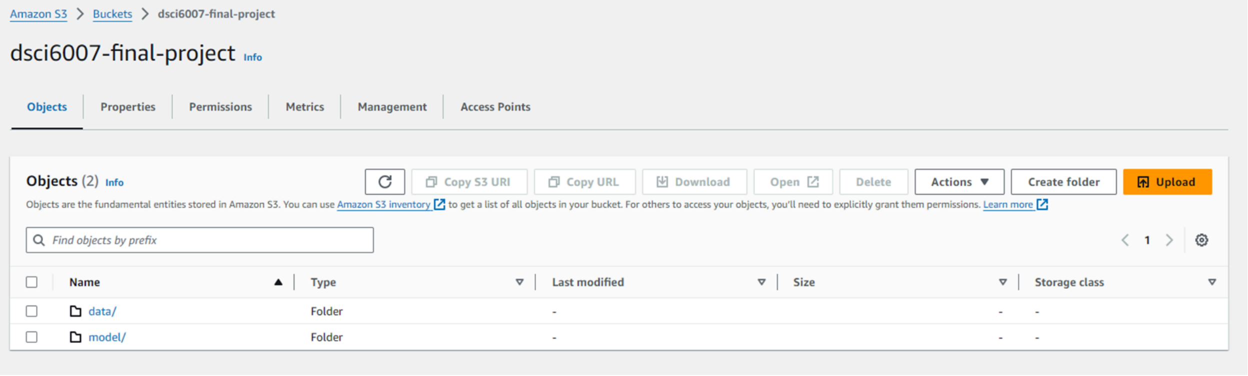
The model was tuned into 100 epochs with a default batch size of 32 and stacking an additional LSTM layer. A GridSearchCV attempt was completed to find the optimal epochs and batch size. The model was prone to overfitting, thus regularization with EarlyStopping and Dropout were used. The final Dropout was set at 50% and added between the two LSTM layers. The EarlyStopping patience was set at 30 epochs. During training, the loss function progression indicated successful learning, with a consistent decrease in loss over time with eventual convergence. With the perturbation effect results, the important features for forecasting Core CPI in this model were past Core CPI, personal consumption expenditure, and M2.

To evaluate the model's performance, model predictions were compared with actual Core CPI values in the test dataset. The root mean squared error (RMSE) for the test data was approximately 1.712, demonstrating the model's accuracy. A visual comparison between the actual and forecasted Core CPI showed that the model's predictions closely aligned with the observed data, indicating that it successfully captured the underlying trends. These results confirm the model's effectiveness in forecasting Core CPI.



Using the trained model, the Core CPI for November 2021 was forecasted and the year-over-year percentage change was calculated. To do this, the model was provided with the most recent twelve months of feature data as inputs to predict the normalized forecast value. The min-max normalization reversal was applied to the data, converting the normalized output from the model to the actual scale of the Core CPI. Using the minimum and maximum Core CPI values from the original dataset, the forecasted Core CPI for November 2021 was calculated. According to the model, the predicted Core CPI for November 2021 is approximately 281.688. For further prediction, the year-over-year percentage change in Core CPI was calculated. Comparing the forecasted value to the Core CPI from 12 months prior, a 4.4% increase was found, suggesting that the forecasted inflation rate for November 2021 would be 4.4%. The combination of accurate predictions and feature importance analysis demonstrates the robustness of our approach and its potential application in economic forecasting. With the modeling portion of the project complete, the model was set to be deployed for user interaction.

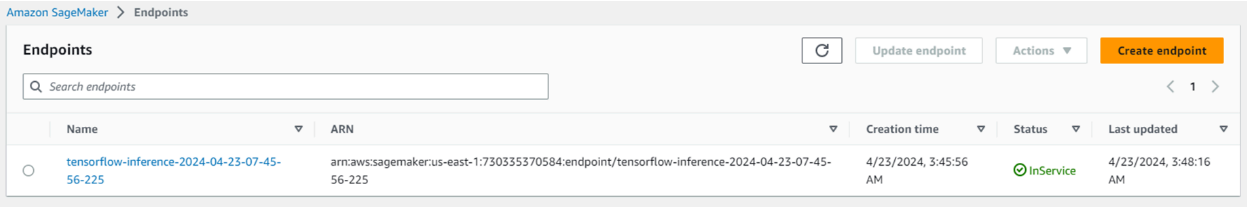
The model deployment process began with saving the model in a TensorFlow SavedModel format. This allowed for the easiest method of deployment on the AWS Sagemaker environment. An archive file with all folders and files relevant to the saved model was created. Special attention was put on the structure of the archive file, as each file or folder needed to be under a folder labeled “1” which was left in the root directory. The model was then able to be uploaded to the same S3 bucket that was utilized earlier for the data, however, the model was stored in a separate directory for ease of access.



Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name ▲	Type	Last modified	Size	Storage class
□	data/	Folder	-	-	-
□	model/	Folder	-	-	-

The all-important archive file path was passed to the “TensorFlowModel” object to be deployed as an endpoint predictor. Minutes of processing time went by, but the endpoint was successfully created and serviceable for invocation. Through the use of previously saved input data for the November 2020 to October 2021 time period and the minimum and maximum Core CPI value across that period, a prediction value was generated from the model endpoint. The prediction was then extracted and reformatted, following a familiar process. The manual unscaling of the prediction value yielded the Core CPI value for November 2021 to be 281.625 and the year-over-year change in the Core CPI value to be 4.38%.



	Name	ARN	Creation time	Status	Last updated
○	tensorflow-inference-2024-04-23-07-45-56-225	arn:aws:sagemaker:us-east-1:730335370584:endpoint/tensorflow-inference-2024-04-23-07-45-56-225	4/23/2024, 3:45:56 AM	✓ InService	4/23/2024, 3:48:16 AM

Conclusion:

In this project, we embarked on a journey to design and implement a robust data engineering pipeline, culminating in the deployment of a machine-learning model for real-world application in economic forecasting. Through meticulous execution and leveraging the capabilities of AWS Sagemaker, we addressed each stage of the pipeline with precision and ingenuity. The initial tasks of data ingestion and storage laid a solid foundation for subsequent analysis, ensuring that the necessary data was readily accessible for processing. With the data prepared and cleaned, we embarked on a journey of exploration, uncovering insights into economic trends and relationships through comprehensive data analysis and visualization techniques. With the model trained and validated, we transitioned to the deployment phase, ensuring that our insights could be readily accessed and utilized. By deploying the model as an endpoint on AWS Sagemaker, we enabled seamless interaction, allowing users to obtain real-time predictions of the Core CPI and its year-over-year percentage change. The culmination of this project underscores the importance of robust data engineering pipelines in deriving actionable insights. Our approach, combining meticulous data analysis, model development, and deployment, offers a blueprint for organizations seeking to harness the power of data to drive informed decision-making and deliver value to stakeholders.

As we reflect on the project, we recognize the opportunities and challenges inherent in such endeavors. Moving forward, we envision further refinements and enhancements to our pipeline, embracing advancements in technology and methodologies to continually push the boundaries of what's possible in the realm of data-driven insights. With the thought of the long-term capabilities of data engineering in mind, we pushed to create a pipeline built for scalability. We hope to further scale this project, using even more AWS services to help us along the way.