# LALLY SCHOOL OF MANAGEMENT

**MGMT 6962: Artificial Intelligence and Machine Learning for Finance**

**Project Two: Supervised Learning**

**Submitted By:**

**Jack Deegan, Clarence Chen, John Lin**

May 1st 2022

## Table of Contents

# AI Project Two: US/Euro Exchange Rate Model Interpretability

## Abstract

Machine learning has become an indispensable tool for the financial industry as it can be used for an inestimable amount of tasks. A few examples of its usages are in the fields of data analytics, marketing, fraud detection and investing. The primary concern moving forward for machine learning's further implementation into the banking industry is in its interpretability. The interpretability of a model is vital for not only the institution but for the customer as well. Knowing how a model reaches its predictions is important because it provides clarity to the reliability, stability and fairness of the model. In the previous project our group produced a long short term memory neural network that was designed with the purpose of predicting the changes in the USD/EUR exchange rate. Our team followed 5 different interpretability methodologies to provide explanations for our models predictions and gain insight on the validity of the model. This paper will provide the reader with an understanding of the importance of interpretability in machine learning, a general understanding of the methods used, and a report on how the LSTM model made its predictions.

**Importance of Interpretability in Machine Learning**

Amongst the largest limitations of machine learning is the lack of explainability of the models predictions which is especially apparent in neural networks. Model Interpretability is extremely important for a number of reasons both for the model user and the people affected by the models predictions. In the case of credit score modeling it is extremely important to be able to access the reasons that a specific credit rating was given because of how great of an impact it has on the customers livelihood. If the model is left unexamined and without explanation, the model may discriminate against protected classes. This provides a substantial legal risk to the firms or banks utilizing the models themselves. Furthermore it poses severe model risk, due to the fact that if the users of the model cannot understand the model they will be unaware of its shortcomings and weaknesses.

The aforementioned necessity to provide explainability for machine learning models is yet more important for neural networks due to their sheer complexity. Simple regression models provide a much clearer insight into the decision making of a model which is a large contributor to its widespread utilization in the financial community. Yet AI is becoming increasingly more prevalent; In a study conducted by McKinsey & Co they found that: "*AI adoption is continuing its steady rise: 56% of all respondents report AI adoption in at least one function, up from 50% in 202 ... Nearly two-thirds say their companies' investment in AI will continue to increase over the next three years*"(Heires 1). It is apparent that AI will be an important tool that financial firms will utilize in the future. Artificial Intelligence can serve many purposes in the financial industry, currently most prevalent of which is fraud detection. A survey produced by the Association of Anti Fraud Examiners found that 31 percent of banking institutions are using AI to detect fraud in 2022 which is nearly double the figure it had been in 2019 (Heire 2).

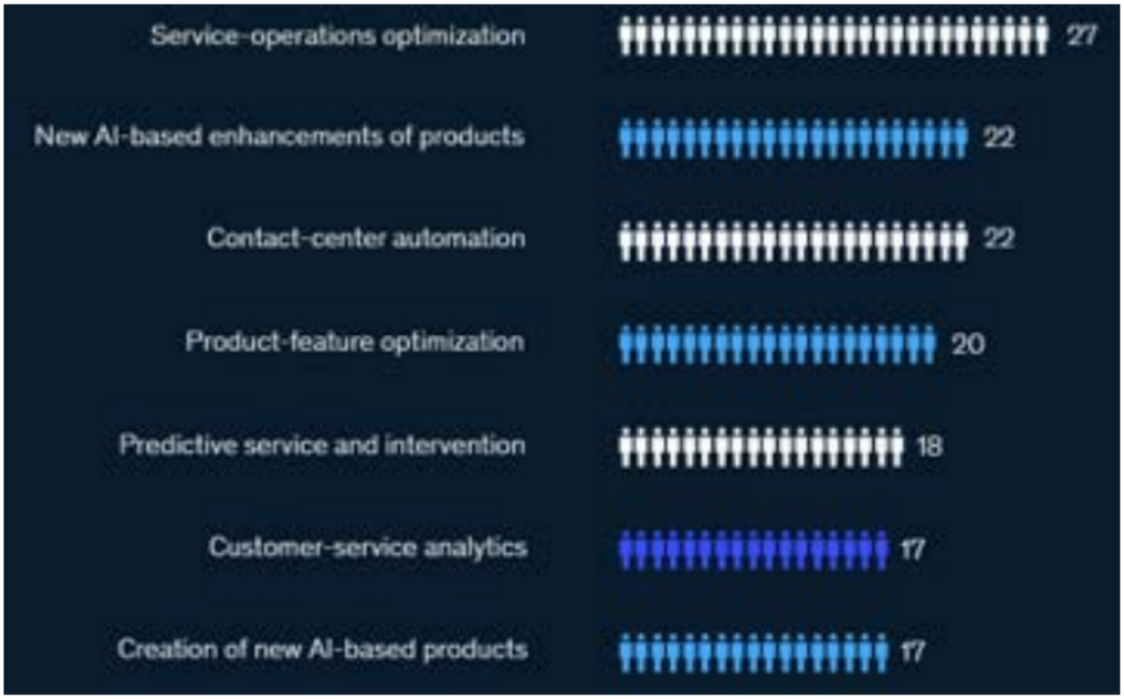


*Figure 1: 2022 McKinsey AI Survey Utilization Results (by Percentage)*

Observing from *figure 1* the most prevalent utilization of AI in the banking sector lies within two primary fields: risk management and backside analytics. The financial industry is very aware of the capabilities of artificial intelligence hence their current drive to incorporate them into their companies infrastructure.

As explained, the fundamental problem faced by AI is in its explainability, thus it has been a primary goal of machine learning scientists to develop their ability to do just that. The “Black Box” problem that AI is faced with is not an easy fix due to its substantial difference in complexity with more orthodox models. Anthony Mancuso a risk modeler from SAS stated that the: *“challenges around explaining AI and ML models, and this feeds into regulatory concerns….Explainability concerns, especially with so-called black box models that regulators scrutinize and may reject, deter many from pursuing or experimenting with AI”(Heires 3).* Because of the black box issue many within the banking industry are waiting for financial regulators to determine clearer legislation on where AI should and should not be used. Others are actively developing their own interpretability methodologies to get ahead of the curve. In this paper we will be using a number of methods to interpret the results of our long short term memory neural network.

## Methodology Description

### *Partial Dependence Plots*

This methodology commonly referred to as PDP plots is a common methodology used by machine learning scientists to gain understanding of how input features impact their model's prediction. In effect the plot shows the effect given a range of input variables upon the dependent variable, and furthermore if that relationship is linear, exponential or complex. For simple regression models a pdp plot should return a linear output, however for complex models the relationship will tend to become more complex. The primary downside to using this methodology is that it contains an inherent bias to highly correlated values. This was especially concerning to our model because many of our features were highly linked.

$$\hat{f}_{x_S}(x_S) = E_{x_C}[\hat{f}(x_S, x_C) = \int \hat{f}(x_S, x_C)dP(x_C)$$

*Figure 2: Partial Dependence Function*

$$\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$$

*Figure 3: Monte Carlo Partial Function for Samples*

### *Individual Conditional Explanation Plots*

Similar to a partial dependence plot, ICE plots measure the effect of an input feature on a model's prediction however on a case by case basis. Both ICE and PDP plots are “model agnostic” which simply means that their applicability is not catered to any specific model rather are applicable to any model. The main delineation is that the ICE Plots measure local effects rather than the average effect, otherwise these can be seen as two different variations of the same method. They also share the same shortcomings, ICE and PDP plots connect unrealistic combinations of features that probabilistically would not exist simultaneously due to correlation.

$$\hat{f}(x) = \hat{f}(x_S, x_C) = g(x_S) + h(x_C), \quad \text{with} \quad \frac{\delta \hat{f}(x)}{\delta x_S} = g'(x_S)$$

*Figure 4: Derivative ICE Plot Function*

### *Accumulated Local Effect Plots*

ALE plots are designed to counter the shortcomings of ICE and partial dependence plots in that they are deemed to be a more unbiased method of gauging a features effect on a dependent variable. It is a global method, meaning that similar to a PDP plot it gauges the average rather than the individual effect on the predictions. The main delineation is that Accumulated local effect plot’s gauge the effect of an input feature given specified ranges in the data rather than continuously. This will allow for usage of models with higher correlated features providing a good representation for our project.

$$\begin{aligned}\hat{f}_{S,ALE}(x_S) &= \int_{z_{0,S}}^{x_S} E_{X_C|X_S=x_S}\left[\hat{f}^S(X_s, X_c)|X_S = z_S\right] dz_S - \text{constant} \\ &= \int_{z_{0,S}}^{x_S} (\int_{x_C} \hat{f}^S(z_s, X_c) d\mathbb{P}(X_C|X_S = z_S)d)dz_S - \text{constant}\end{aligned}$$

*Figure 5: Accumulated Local Effect Function*

### *Permutation Importance Plots*

PIP plots take a different approach to accessing a features impact on a model, rather than accessing the value effect on the prediction PIP plots plot the change in prediction error given a permutation of a feature. Permutation importance plots provide a global outlook into the models predictions and fully takes into account features interactions with all other variables. The plot is in the form of a column chart, the largest size columns provide the largest amount of error on the models predictions and thus are the most important to the model. Similar to PDP and ICE plots, PIP plots suffer from poor performance with models containing high correlation as some permutation combinations would be unrealistic.

### *Local Interpretable Model-Agnostic Explanations*

The LIME method seeks to explain a singular prediction by creating a localized model based upon a select sample of the dataset. The model created off of the specified point is often a simplification of the original model; two examples are ridge regressions and decision trees. In fact it is an objective of the function to minimize the model's complexity parameter denoted by omega in figure 6. The purpose of LIME is to provide the user with a real understanding on a case to case basis why a model made the prediction it did. This is especially vital for models deemed to be “black boxes” because it provides an understanding of a complex model's predictions like a simpler model could. For neural networks which are at the center of the black box conversation it is a particularly good tool to utilize.

$$\psi(x) = arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

*Figure 6*: Local Surrogate Model Function

## Model Data Description

The data used consisted of various features based on what we thought would be useful in predicting the US to Euro Foreign exchange rate. For example, we included fossil fuel costs as we believed they would provide a good estimation of the buying power of US currency, due to how important oil is to the economy. In addition, we included several different features that estimate various commercial papers to further refine our measures on the economy. We also included stocks measures such as the SP500 and the Dow Jones Composite average, both of which were included to compare it with market effects. United States treasury security rates were added as they are directly related to the value of the dollar. As rates of securities go up the value of the dollar rises in unison and inversely to the contrary. Finally, we included GDP, which posed issues merging with the other data as the data was quarterly, while the rest of the data was daily, and thus we used linear interpolation on the GDP data points to make it “daily”. The first step in aggregating this data was to make all date data be formatted in the same manner, to let all of the data be matched by indexing the date. Afterwards, all of the data sans the date (which has already been converted to the index) is converted to numeric format, to allow for mathematical operations to be performed on it, as well as filling non numeric data with NaN instead, making them easy to remove en-masse. Finally, instead of just predicting the exchange rate directly, we calculate the log return rate and predict that instead as it produces far more stationary data ensuring that our time series model is actually predicting useful data.

## Model Description

The model we analyzed was a Long Short-Term Memory (LSTM) machine learning model. LSTM was used as it is a model that is suited for analysis of time series data. The model was built on a sequential model from the keras package of python, using 2 LSTM layers and 2 corresponding Dropout layers. The double layer of LSTM allows for a more complex modeling of the patterns in the data presented, while the dropout layers were present to reduce overfitting in the data. As part of an earlier project, we iterated through various neuron counts along with various epoch and batch sizes to try and find an "optimal" set of parameters that resulted in the highest r square value for out of sample testing results.

There was considerable difficulty in adapting this model to work with the different interpretability metrics mentioned, including some sacrifice in accuracy, as the LSTM models takes in 3 dimensional data in the shape of ( number of samples, time series length, features ), while most interpretability metrics work based on 2 dimensional data. Thus, we had to set the time series length to 1 to allow for an active reshaping layer for the data while inside the LSTM model itself, but as previously mentioned, this sacrificed some accuracy in the model as it stopped us from being able to consider longer time series lengths for this analysis.

## Results

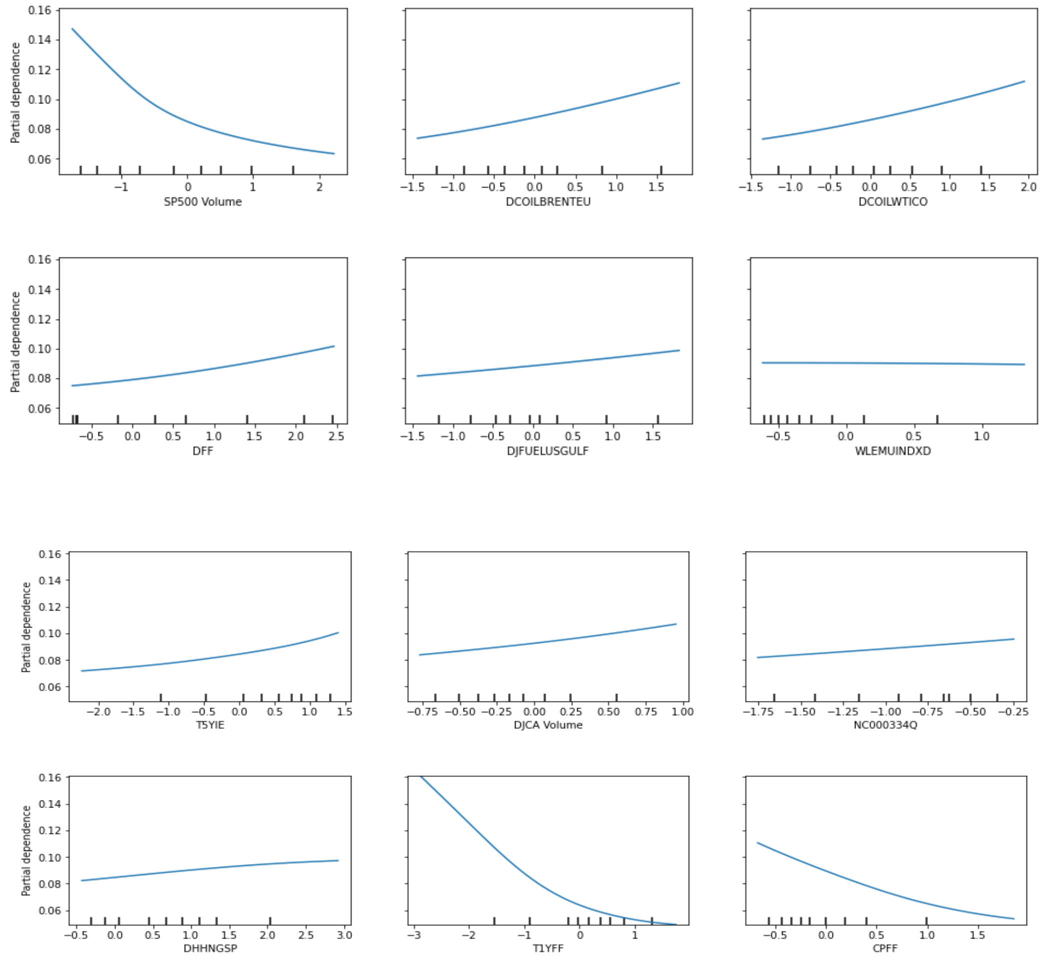
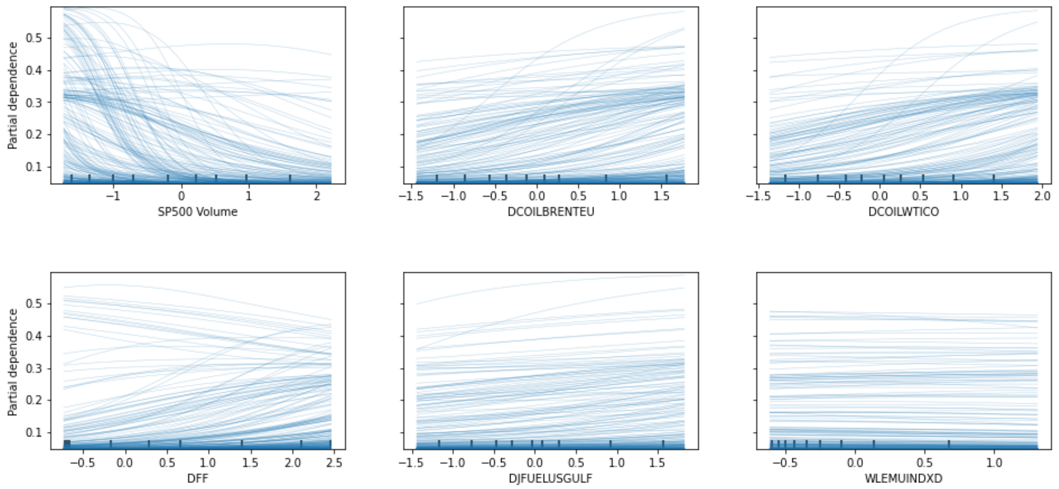


*Figure 7: PDP Plots*



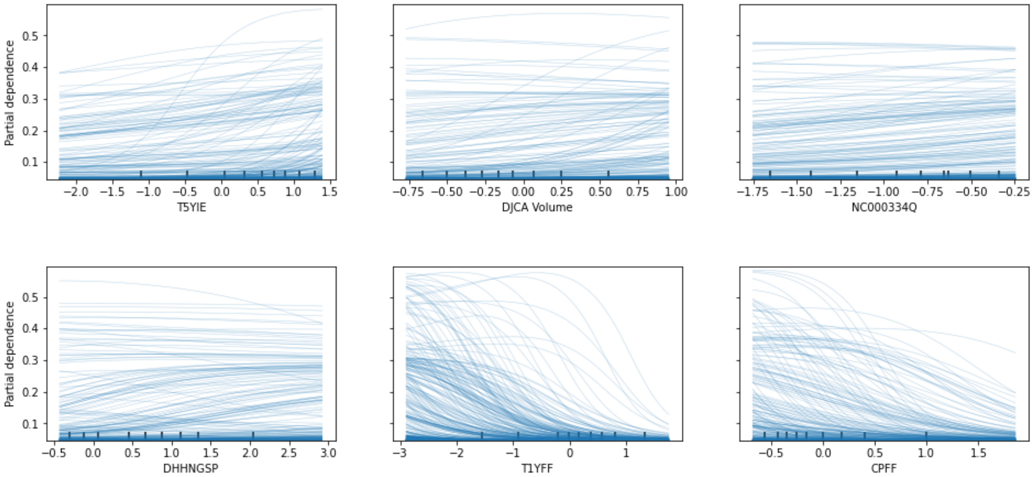
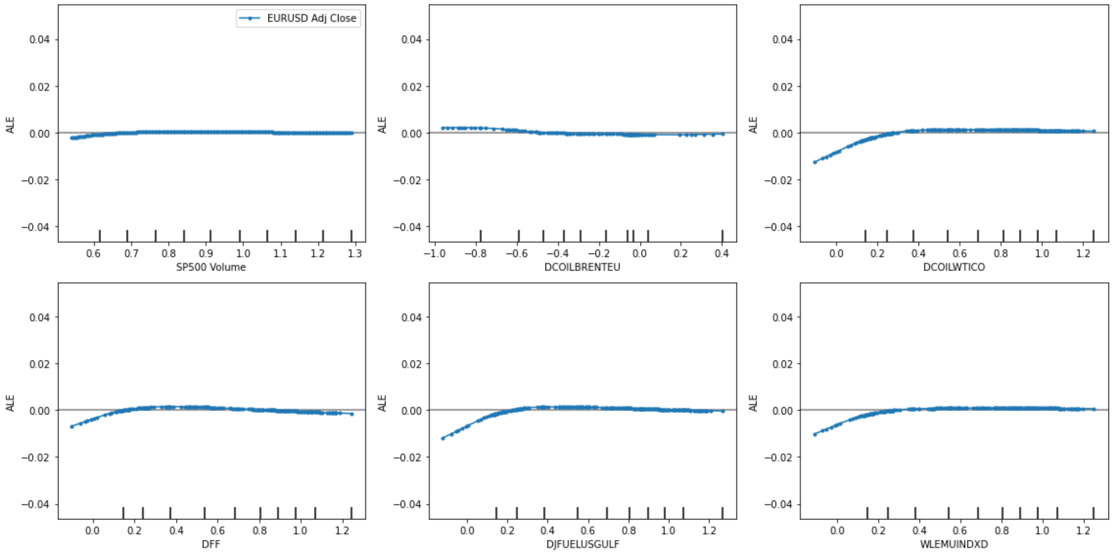


*Figure 8: ICE Plots*

From the ICE plots above, which were made with 400 subplots, we can see that while there is significant heterogeneity within some of the plots, such as in SP500 volume, most of them still follow the general overall relationships we saw previously. In the charts with high levels of heterogeneity, it suggests that there are likely correlations in those features with other ones. For example, for SP500 volume which noticeably has 2 very distinct patterns in its subsamples, which may potentially be caused by correlation with the other SP500 measures, as they are likely to be closely related.



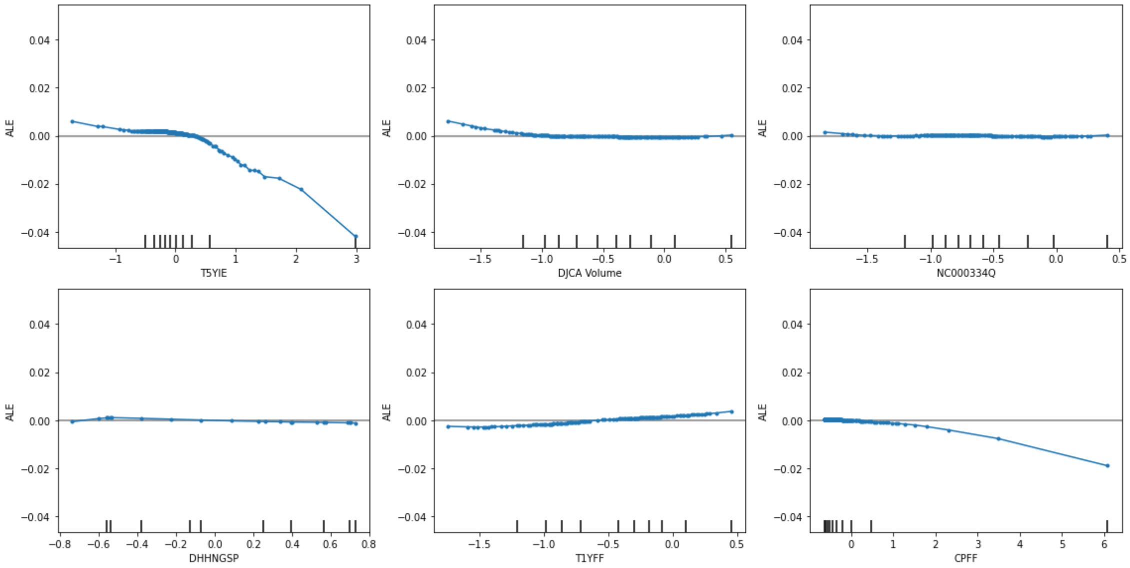


*Figure 9: ALE Plots*

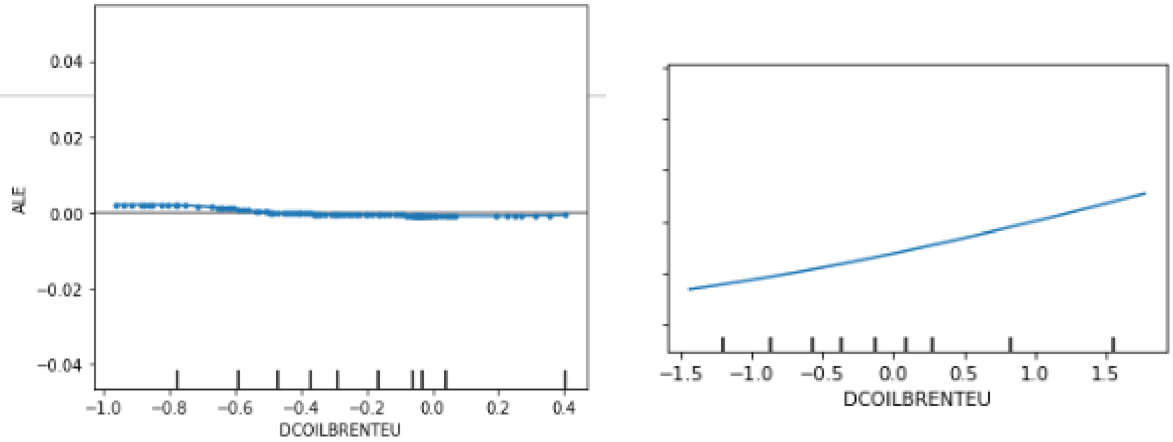
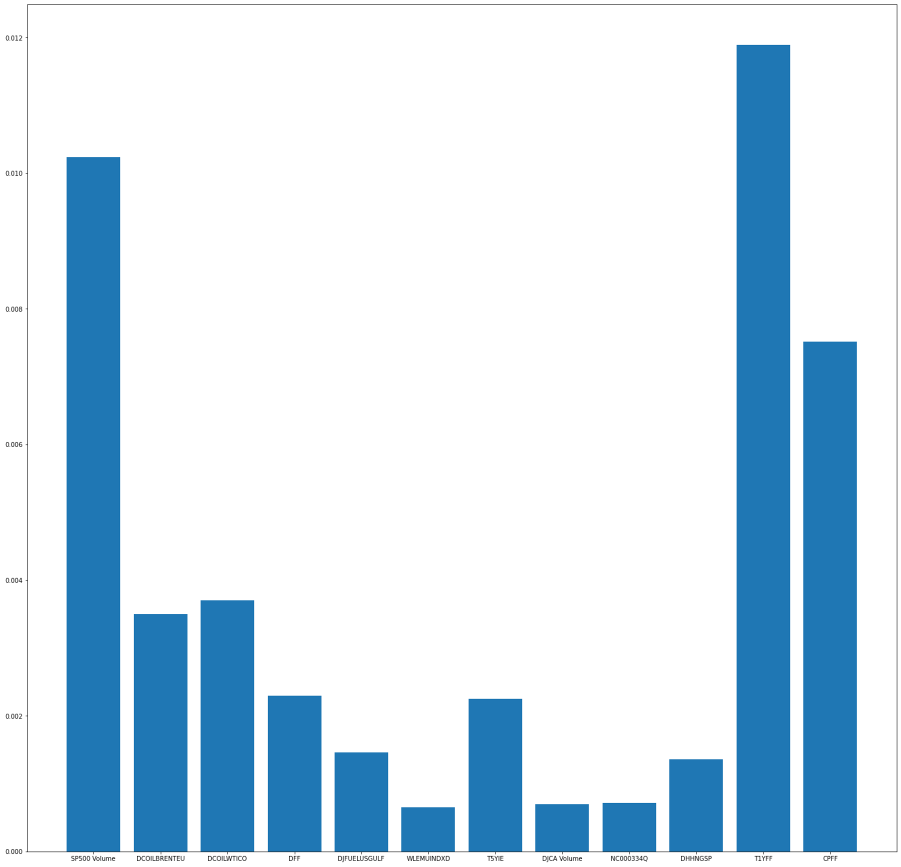


*Figure 10: ALE vs. PDP Comparison*

As shown in Figure 4, we can see that the ALE plots have a common pattern of low correlation to the target variable. This differs greatly from the previous ICE and PDP plots which can be explained due to ALE plots strength in providing an unbiased measure of feature and target relationships. This relationship is directly visualized in Figure 5, showing how there is a large difference in the generated slopes of DCOILBRENTEU with the ALE plot providing a much more flat slope while the PDP plot shows correlation with a noticeable slope. The majority of the ALE plots are flat with very little variance except for the absolute extremas of the graphs. However, certain features such as T5YIE(5-Year Breakeven Inflation Rate) provide a different outlook with a noticeable variable relationship with the target variable. These anomalies could potentially be used to create a new dataset with variables that actually show high correlation to the target variable and may lead to a more accurate model.



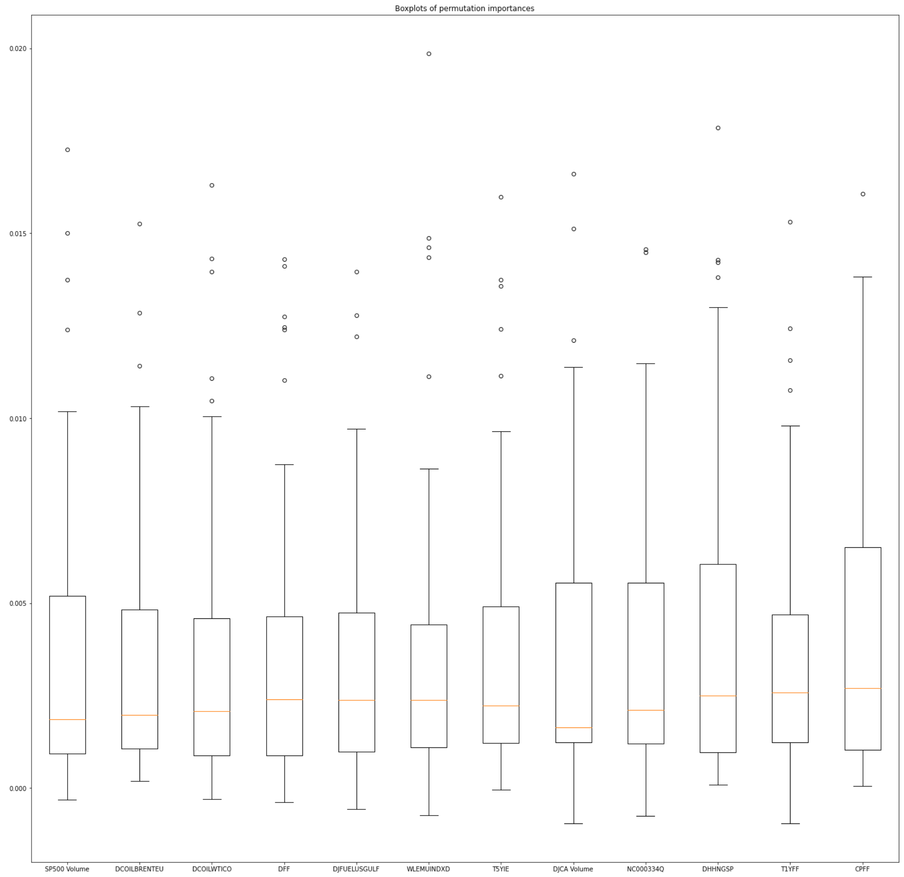
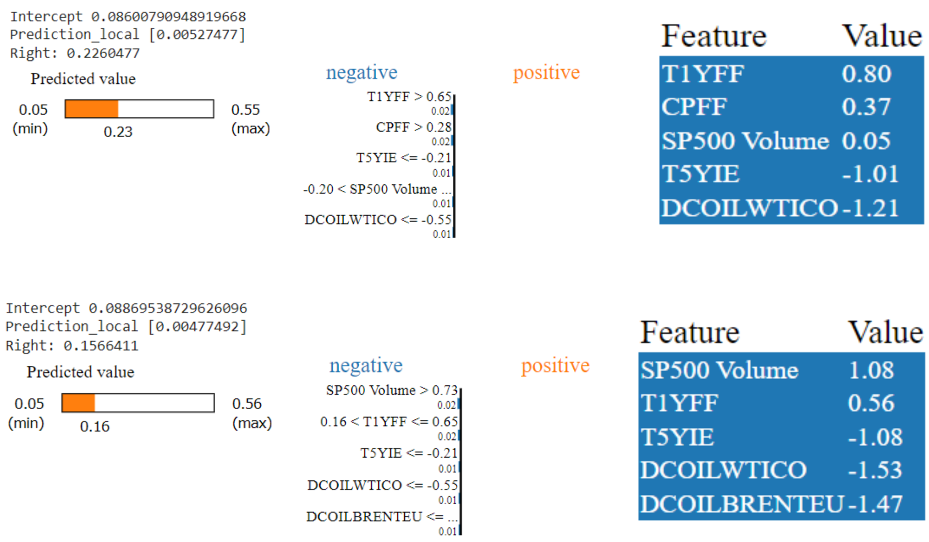


*Figure 11: Permutation Importance Plots*

Viewing the PIP plots shown in Figure 7 show three highly correlated variables, S&P 500 volume, 5-Year Breakeven Inflation Rate, and the Federal Funds Effective Rate. These results show that our model has an overdependence in certain variables. This issue shows a lack of model flexibility in our LSTM solution. Without these specific variables, our model would perform even worse than it already has. However, these results are not fully unbiased. PIP plots are mainly dependent on the errors of the model itself. As we have established, our model isn't the most fitting model for solving this solution and determining errors in our model may not lead to any conclusive results. PIP plots would be better for a model that performs well in order to tweak the model for better accuracy and would potentially be better for an improved form of our model.



```
Intercept 0.08600790948919668
Prediction_local [0.00527477]
Right: 0.2260477
```

| Feature | Value |
|---|---|
| T1YFF | 0.80 |
| CPFF | 0.37 |
| SP500 Volume | 0.05 |
| T5YIE | -1.01 |
| DCOILWTICO | -1.21 |

```
Intercept 0.08869538729626096
Prediction_local [0.00477492]
Right: 0.1566411
```

| Feature | Value |
|---|---|
| SP500 Volume | 1.08 |
| T1YFF | 0.56 |
| T5YIE | -1.08 |
| DCOILWTICO | -1.53 |
| DCOILBRENTEU | -1.47 |

*Figure 12: Random selection of 2 LIME test results*

As we can see from the LIME test results above, we see that there are extremely low weights for most features in the functions being generated by the LIME's local model. However, the highly weighted features are consistent with the results of prior methods such as the PIP plots showing the same top five features. However it is important to note that the local models themselves were to a degree inconsistent as well highlighted from the fact that the weights vary a large amount across predictions. Observing the disparity in prediction from the local LIME model and the actual, we can infer that our model lacks predictability, although this may simply be due to the sheer difficulty in even predicting foreign exchange, not just due to the poor fit of our model on the data.

## Conclusion

The five interpretability methodologies utilized to analyze the team's long short term memory neural network brought to light fascinating differences in each of the methodologies as well as the validity of the model. In the creation and observation of the plots we saw first hand the shortcomings of each method. Most notably how the ALE plot differed from the PDP and ICE plots, due to an overabundance of correlated data within our feature set. The ALE plot did not show the profound movements on either extreme as the partial dependence plots had because it had not combined unlikely combinations of the two features. Gaining insight on the model it was found from the PIP plot that the S & P 500 volume, 5-Year Breakeven Inflation Rate, and the Federal Funds Effective Rate were the most decisive in determining the exchange rate within our model. However a common trend throughout was the hints of unreliability of the underlying model, for example in almost all of the ALE plots the features used provided little to no predictive power on the dependent variable. Overall our group learned a significant amount about the LSTM but also about the importance of interpretability in machine learning and the methods used to accomplish it.

## References

*Heires, K. (2022, March 25). Artificial Intelligence and Machine Learning Grow in Financial Services – with Caveats. Katherine Heires.*

*Molnar, C. (2022). Interpretable Machine Learning:*

*A Guide for Making Black Box Models Explainable (2nd ed.).*

*christophm.github.io/interpretable-ml-book/*