Capstone Public Project Summary

# Forex Spread Data Analysis

Nicolas Fernandez

**Academic Program:** Economics and Business

**University:** Central European University

In partial fulfilment of the requirements for the degree of Master of Sciences in Business Analytics

**Supervisor:** Zoltan Toth

**Place of Submission:** Vienna, Austria

**Date/Year of Submission:** June 10, 2024

# Table of Contents

## Introduction

In the highly competitive world of foreign exchange (forex) trading, the spread – the difference between the bid and ask prices – is a critical factor that can significantly impact trading costs and profitability. Spreads are influenced by various factors, including market volatility, liquidity, and the specific broker's policies. Despite their importance, there remains a lack of transparency in how brokers report and manage spreads across different volumes and trading instruments. This report details the analysis of spread data provided by various brokers to uncover patterns, inconsistencies, and optimal choices for traders under different trading conditions. By employing comprehensive visualizations and predictive models, this analysis provides insights for traders aiming to optimize their broker selection based on their trading volume and preferred instruments.

## Methodology

The data for the analysis was pulled from two separate APIs through which python scripts were created to access, specifically MetaTrader and cTrader FIX Protocol APIs. The scripts were developed to pull depth of market data periodically for user-specified symbols. For the purposes of this analysis the two symbols that were chosen were the following:

- EURUSD – Euro/US Dollar currency pair
- XAUUSD – Gold/US Dollar metal/currency pair

The depth of market data that was pulled contains timestamps of when the data was accessed, broker names, the foreign exchange symbol, trading volume sizes, ask and bid prices for the respective volume, if available and the API source from which the data was pulled. From the collected data the spread data was calculated using the ask and bid prices for each respective volume of trade. The volume column was converted into lot sizes to make the lots the same irrespective of symbol.

The data was organized by grouping brokers, symbols, and trading volume sizes and aggregating the respective means of ask prices, bid prices, and spread values. This allowed for viewing the data in a more meaningful way for making spread comparisons between brokers. Additionally, timestamp data was formatted to the minute rather than the second to allow for easier matching of ask prices to bid prices within the data.

**Exploratory Data Analysis**

The distributions of the log-transformed spread values were evaluated and can be seen from the figure in Appendix A. It shows a normal distribution overall in the spread values for both symbols indicating that the variance in the data is stable when viewed by their log values. Given that the distributions approach normal, no additional transformations of the spread data are necessary.

While spread data at specific volumes of trading are reported by brokers, a low spread at a specific volume does not necessarily mean that the specific broker reporting that spread has the best spreads overall for that volume of trading. This difference amounts to the nature in which lots are purchased and sold and how these transactions are calculated. The plot in Appendix B illustrates this. FX Pro has the largest spread in EURUSD while Pepperstone has the second lowest spread in this metric. From other visualizations Pepperstone had performed worse when examining overall spreads however this was likely due to Pepperstone offering trades at higher volumes on average than other brokers within the data, which also typically come with higher spreads. Conversely, Pepperstone performs the worst for XAUUSD when calculating for 10 lots. It is likely that certain brokers excel more with currencies whereas others perform better with instruments of other types and/or natures.

**Models for Predicting Spread Values**

Several models were created to create accurate predictions of spread values from the data, the results for which can be seen in Appendix C. The target variable selected was the log-

transformed spread value with root mean squared error (RMSE) as the loss function for evaluations. From the RMSE results from each symbol, the Random Forest model performed the best with a very low RMSE value. This suggests that the general relationship between spread data and the explanatory variables is non-linear and therefore best represented by a machine learning model that can account for non-linearity through methods of regularization. Taking a step back, however, the RMSE values in general are very low. The data quality that is available is likely poor. No amount of modeling can correct this issue if it exists with the only fix being accumulating more data. From general intuition, however, it's likely that there is a strong linear relationship between spread data and lot/volume.

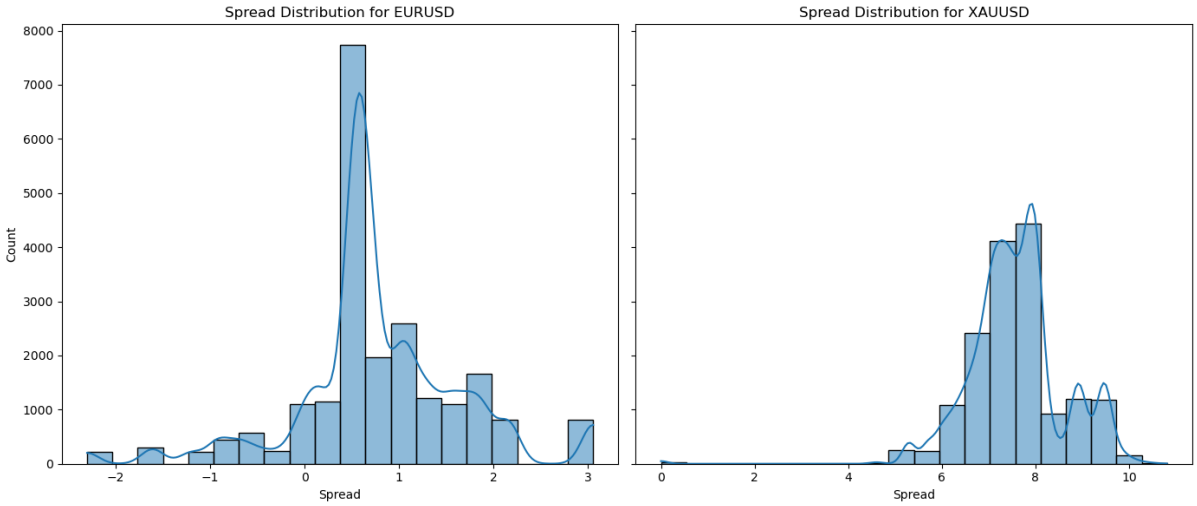
**Overall Analysis and Conclusion**

From examining the data, it's clear that there are differences in spread data amongst brokers. Brokers tend to report close to 0 spreads at different volumes within their depth of market data to entice traders to employ their services. What has always been nebulous has been whether these reported spreads are accurate and consistent. With the inferences made from the exploratory data analysis, it is evident there are differences in the spreads across brokers and that the volumes being traded at are significant.

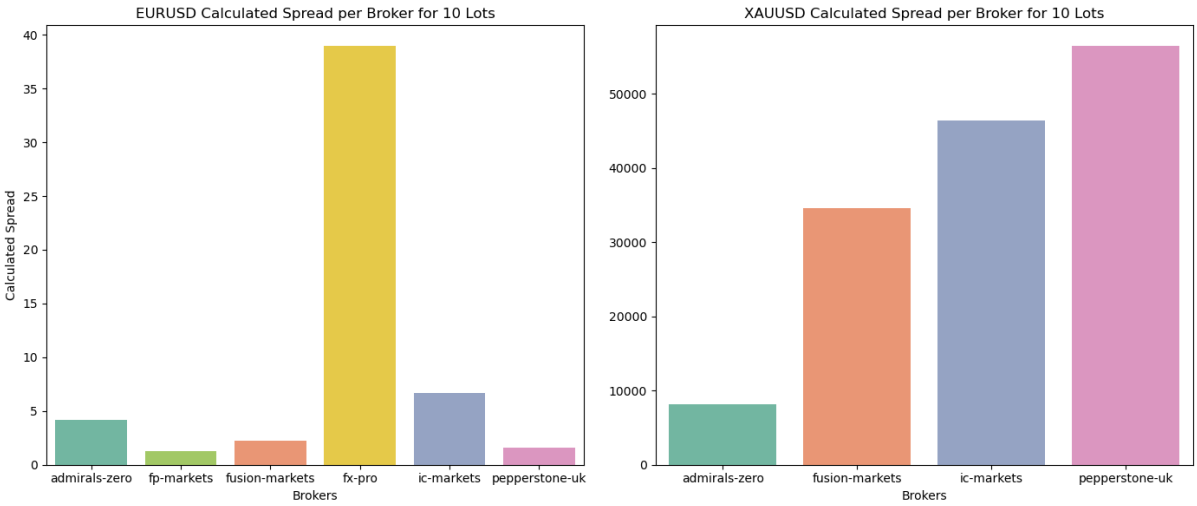
These types of discrepancies are not usually disclosed by brokers and trading platforms and are typically left to the user to research. This analysis highlights the importance of conducting detailed research and utilizing predictive models to make informed decisions in the highly competitive and often opaque forex market. For that purpose, this analysis hopes to provide inferences for a tool for ongoing analysis to be made.

## Appendices

Appendix A:




Appendix B:




Appendix C:

| | EUR Model | Train RMSE | Test RMSE |
|---|---|---|---|
| 0 | Benchmark | 0.893559 | 0.895417 |
| 1 | OLS Base | 0.380516 | 0.375682 |
| 2 | OLS All | 0.276958 | 0.275442 |
| 3 | RF | 0.051764 | 0.135188 |
| 4 | GBM | 0.174478 | 0.178445 |

| | XAU Model | Train RMSE | Test RMSE |
|---|---|---|---|
| 0 | Benchmark | 1.006980 | 1.035854 |
| 1 | OLS Base | 0.537452 | 0.590858 |
| 2 | OLS All | 0.493528 | 0.544425 |
| 3 | RF | 0.154607 | 0.418624 |
| 4 | GBM | 0.474441 | 0.536514 |