ALMASTONE - CASE STUDY

Financial Analysis & Tech Exploration

DANIEL POOLE

TABLE OF CONTENTS

PART I: FINANCIAL ANALYSIS

Access financial data through public APIs, aggregate into a central repository and use python to analyze.

Background Research - SLC Agricola

- One of Brazil's largest soybean, corn and cotton producers
- Own 23 farms that produce commodities, seeds & cattle
- Products sold in Brazil & abroad
- Revenue via trading commodities on international exchanges (quoted in USD)
- Revenue actively exposed to variations in commodity prices & FX Rate
- Hedge through forward contracts, futures & options

[QUESTION #1 - Download financial data for SLC Agricola using a python script]

Stock price & commodity futures data from the last (5) years was downloaded from Yahoo Finance via API using the "yfinance" library. Quarterly income statement data for SLC Agricola was downloaded as a CSV file from Yahoo Finance and loaded into a dataframe using the pandas library.

[QUESTION #2 - Explore relationship between commodity futures, FX rate & stock price]

[**2.1**] Plot SLC Stock Price vs Commodity Futures & FX Rate

- Direct linear relationship between SLC Stock Price & Commodity Futures
- Inverse relationship between SLC Stock Price & BRL (nonlinear). A strong Real would mean that Brazilian commodities look more expensive to the foreign currency holder (importers). When the BRL is weak, the company has performed its best. This is because Brazil is the world's biggest exporter of beans and a large exporter of corn. A country chooses their import products based on delivery economics (most of the time) and will import from the cheapest offer + freight.
- See **Appendix I** for plots

[**2.2**] Pearson's Correlation Coefficient Analysis

Pearson's correlation coefficient is a measure of the linear correlation between two variables. It ranges from -1 to 1, where a value of 1 indicates a perfect positive correlation (i.e., a direct relationship in which the values of one variable increase as the values of the other variable increase), a value of -1 indicates a perfect negative correlation (i.e., an inverse relationship in which the values of one variable decrease as the values of the other variable increase), and a value of 0 indicates no correlation.

The Pearson's correlation coefficient is calculated as the covariance of the two variables divided by the product of their standard deviations. The formula for the Pearson's correlation coefficient, denoted by the symbol r, is as follows:

$$r = \text{cov}(X, Y) / (\text{std}(X) * \text{std}(Y))$$

Pearson's correlation coefficient is useful for identifying linear relationships between variables, which makes it a common way to evaluate the strength and direction of a relationship between two continuous variables.

Table 1: Pearson's Correlation Analysis for Stock Price vs Commodity Futures & FX Rate

Commodity Futures & FX Rate	Pearson's Correlation Coefficient
Soybeans	0.88
Corn	0.91
Cotton	0.81
BRL FX Rate	-0.46

[**2.3**] Linear Regression Analysis

Linear regression is a statistical method used to model the relationship between a dependent variable (also known as the response variable or output variable) and one or more independent variables (also known as predictor variables or input variables). The goal of linear regression is to find the best-fitting line, or hyperplane in multiple dimensions, that describes the relationship between the dependent and independent variables.

Table 2: Linear Regression Coefficients for Stock Price vs Commodity Futures

Commodity Future Contracts	Linear Regression Coefficient
Soybeans	0.000472
Corn	0.009390
Cotton	0.024469

Table 3: Linear Regression Model Evaluation

Error Metric	Value	% Error
MAE	0.596	9.9%
MSE	0.5746	9.6%
RMSE	0.758	12.7%
R-squared	0.851	14.2%
Mean	5.990	n/a

[**QUESTION #3 - Analyze SLC Agricola financial performance in recent years**]

- Seasonal decreases in gross profit during Brazil's summer months (Dec, Jan, Feb)
- Z-score analysis shows that SLC Agricola has performed above average in its last 2 years of operation relative to its overall performance over the last 5 years.
- Plotting SLC Agricola Gross Profit and Commodity Futures vs Time show strong correlation between company financial performance & commodity futures.
- Unusual deviations in 2020 can be attributed to global macro climate & COVID.
- See **Appendix II** for plots of SLC Agricola Financial Performance vs Time.

PART II: TECH EXPLORATION

Incorporate ML into market research and to create a base design for an ML pipeline

[**QUESTION #1 - Download historical SOFR 3M, 6M and 1Y Term data**]

Historical SOFR Term data downloaded from CME Datamine

[**QUESTION #2 - Create an ML model that forecasts SOFR Term Spreads**]

Secured Overnight Financing Rate (SOFR):

The SOFR is calculated as the volume-weighted median of the rates at which eligible participants in the tri-party repo market, an important segment of the money markets, lend cash overnight to other eligible participants in exchange for Treasury securities. This rate is representative of the collateralized borrowing costs of the broad market

[**2.1**] Import macro economic data via FRED API

10Y-2Y Treasury Yield Spread:

The "10-2 spread," is a measure of the difference in interest rates between long-term and short-term U.S. Treasury bonds. This spread is often used as an indicator of the slope of the yield curve, and it can provide insight into the current state of the economy and the expectations of market participants for future economic conditions.

A positive 10-2 spread (i.e., when the 10-year Treasury yield is higher than the 2-year Treasury yield) is considered normal and indicates that market participants expect the economy to be stable and to grow in the future. In this scenario, investors will demand a higher yield for longer-term bonds in order to compensate for the increased risk of inflation.

On the other hand, a negative 10-2 spread (i.e., when the 2-year Treasury yield is higher than the 10-year Treasury yield) is considered inverted, this indicates that market participants expect the economy to slow down or experience a recession in the future.

When the yield curve is inverted, it suggests that market participants are more bearish about the future economic conditions, and investors will tend to demand a lower yield for longer-term bonds as they are less willing to bear the risk of inflation.

An inverted yield curve could affect the SOFR and lead to a decrease in market demand for short-term interest rate products, such as SOFR-indexed swaps and futures, as investors may anticipate a decrease in short-term interest rates in a recession. This decrease in demand could lead to a decline in the SOFR rate.

Effective Federal Funding Rate (EFFR):
The Effective Federal Funds Rate (EFFR) is a measure of the average interest rate at which banks lend overnight funds to other banks that are members of the Federal Reserve System. The EFFR is a weighted average of the rate on overnight federal funds transactions reported by depository institutions and is intended to be representative of the effective rate of interest paid or earned on such transactions.

The SOFR is also used as a benchmark for interest rate derivatives, such as SOFR-indexed swaps and futures. If the Federal Reserve were to change its monetary policy, the EFFR would likely change, which in turn could affect the market demand for SOFR-indexed derivatives and thus affect SOFR rate.

It's important to note that the EFFR and SOFR are different indicators of the interest rate markets and reflect different aspects of the economy, but they are both key indicators of short-term interest rates and can affect the demand and supply of short-term financial products.

[**2.2**] Create ARIMA ML model that predicts spreads between SOFR Term contracts

AutoRegressive Integrated Moving Average (ARIMA) ML Model:
ARIMA is a class of statistical models for analyzing and forecasting time series data. An ARIMA model is a generalization of an autoregressive (AR) model, which models the dependence between an observation and a number of lagged observations, and a moving average (MA) model, which models the dependence between an observation and the errors from a moving average process.

An ARIMA model is specified by three parameters: p, d, and q. Once these parameters are determined, the model can be used to make forecasts of the time series. The predictions are made using the estimated parameters of the model and the past data of the series.

"**p**" = the number of autoregressive terms. The autoregressive component models the dependence between an observation and a number of lagged observations.

"**d**" = the number of differences needed for stationarity. If the series is not stationary, differences of the observations are taken (d term) until a stationary series is obtained.

"**q**" = the number of moving average terms. The moving average component models the dependence between the observation and an error term from the moving average process.

[**2.3**] Evaluate model & predict SOFR spreads for the next year in monthly intervals

Mean Absolute Error (MAE) and Mean Squared Error (MSE) measure the difference between the predicted values and the true values. When evaluating the performance of a machine learning model, it's often useful to express the Mean Absolute Error (MAE) and Mean Squared Error (MSE) as a proportion of the mean of the target variable. This helps to provide a sense of the relative magnitude of the error and allows for more meaningful comparisons between models.

Table 4: ARIMA Model Performance on Test Dataset

Spread	Mean	MAE	MSE	MAE % Error	MSE % Error
SOFR 1Y-3M	0.558	0.105	0.020	18.8%	3.6%
SOFR 1Y-6M	0.225	0.078	0.009	34.7%	4%
SOFR 6M-3M	0.334	0.055	0.005	16.5%	1.5%

Table 5: Predicted SOFR Term Spreads (1Y Forecast @ Monthly Intervals)

Date	SOFR 1Y-3M	SOFR 1Y-6M	SOFR 6M-3M
2023-01-31	0.556	0.195	0.138
2023-02-28	0.542	0.190	0.125
2023-03-31	0.528	0.180	0.121
2023-04-30	0.521	0.175	0.117
2023-05-31	0.515	0.159	0.130
2023-06-30	0.506	0.140	0.148
2023-07-31	0.496	0.120	0.166
2023-08-31	0.487	0.117	0.164
2023-09-30	0.482	0.099	0.186
2023-10-31	0.474	0.097	0.185
2023-11-30	0.466	0.079	0.205
2023-12-31	0.463	0.066	0.223

See **Appendix III** for plots of SOFR Term Spreads vs Time, including predicted values.

[QUESTION #3 - Strategy for trading SOFR contracts through swaps? Risk vs Return]

Evaluation of the trained ARIMA models show that my model predicting the SOFR6M - SOFR3M spread was most accurate on the test dataset, indicated by lower percent errors for the MAE & MSE. Therefore, my trading strategy would be focused on running swaps between this spread based on predicted values from the associated ARIMA model. The predicted values for this model indicate a potential increase in the spread over the next year. To profit off of this forecast I would take a long position by buying the SOFR6M contract and selling the SOFR3M

contract. Feeding input data for returns, I would calculate the Sharpe ratio to account for Risk vs Return and adjust the ARIMA model accordingly.

Sharpe Ratio:
The Sharpe ratio is a measure of risk-adjusted return, which compares the return of an investment to that of a risk-free asset, such as a government bond, and adjusts for the volatility of returns. It is defined as the excess return (or risk premium) of the investment over the risk-free rate, divided by the standard deviation of the investment's returns.

A higher Sharpe ratio indicates that the investment's returns are more consistent relative to the level of risk taken. A negative Sharpe ratio indicates that the strategy has underperformed the risk-free rate. It's widely used to evaluate the performance of investment portfolios, and usually a higher ratio is better, as it means that the portfolio has higher returns for the same level of risk.

[QUESTION #4 - Create a diagram that shows how you would automate the ML pipeline]

Data sourcing and aggregation:
This step involves collecting and combining data from various sources, such as databases, CSV files, and APIs, into a single location, such as a data lake or a data warehouse. The data should then be cleaned and transformed to ensure it is in the right format for the ML algorithm.

Data storage:
The cleaned and transformed data should then be stored in a secure location that is easily accessible to the ML algorithm. This could be a database, a file system, or a cloud-based storage service, such as Amazon S3 or Google Cloud Storage.

ML model development:
In this step, the ML algorithm is trained using the collected and cleaned data. This process involves selecting an appropriate algorithm, tuning its parameters, and evaluating its performance.

Model hosting:

Once the ML model is trained, it should be deployed to a hosting environment, such as a cloud-based platform, such as Amazon SageMaker, Google Cloud ML Engine, or Microsoft Azure ML. The deployed model should be accessible via an API endpoint, which allows external systems to send data to the model and receive predictions in return.

Model management:

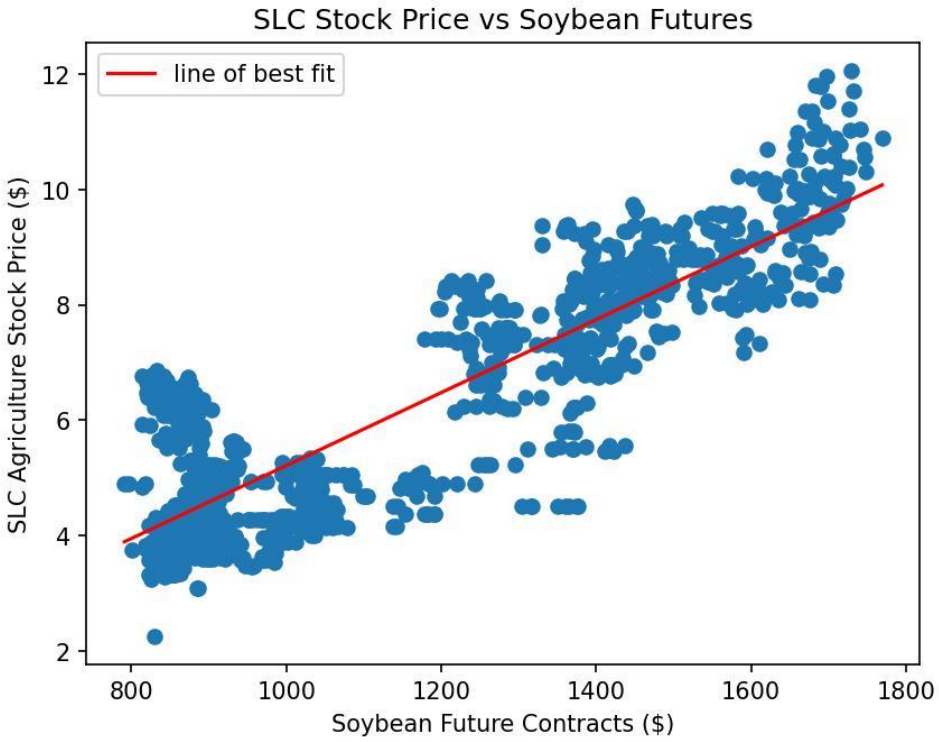
To ensure the ML model remains accurate over time, it may need to be periodically retrained or fine-tuned. This process of keeping track of models, updating, or replacement of a stale model with a new one should be fully automated.

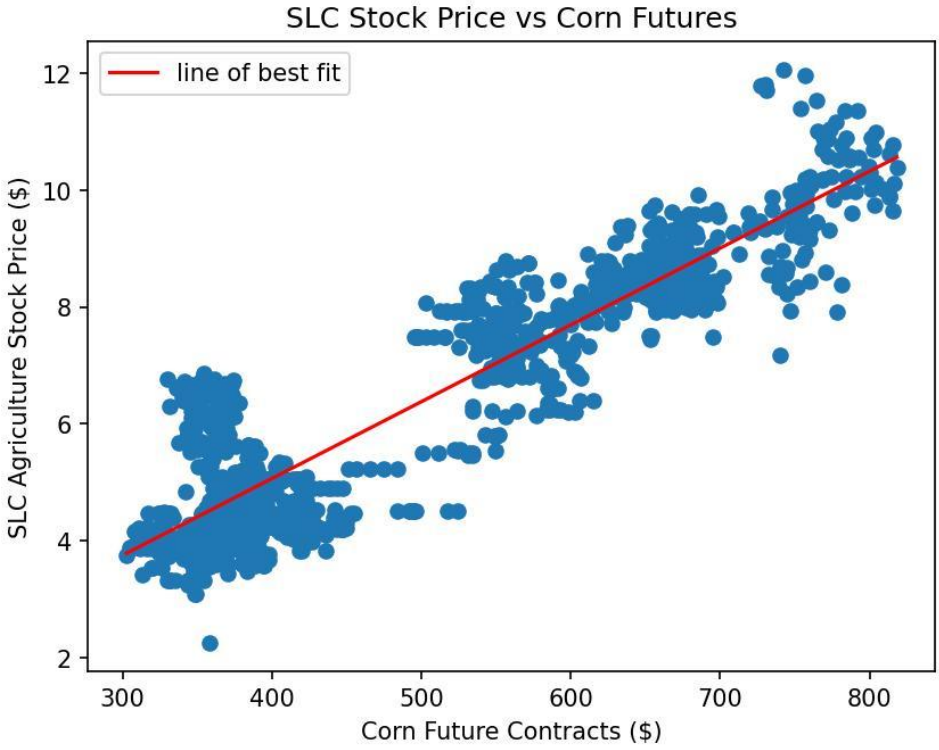
Monitoring and logging:

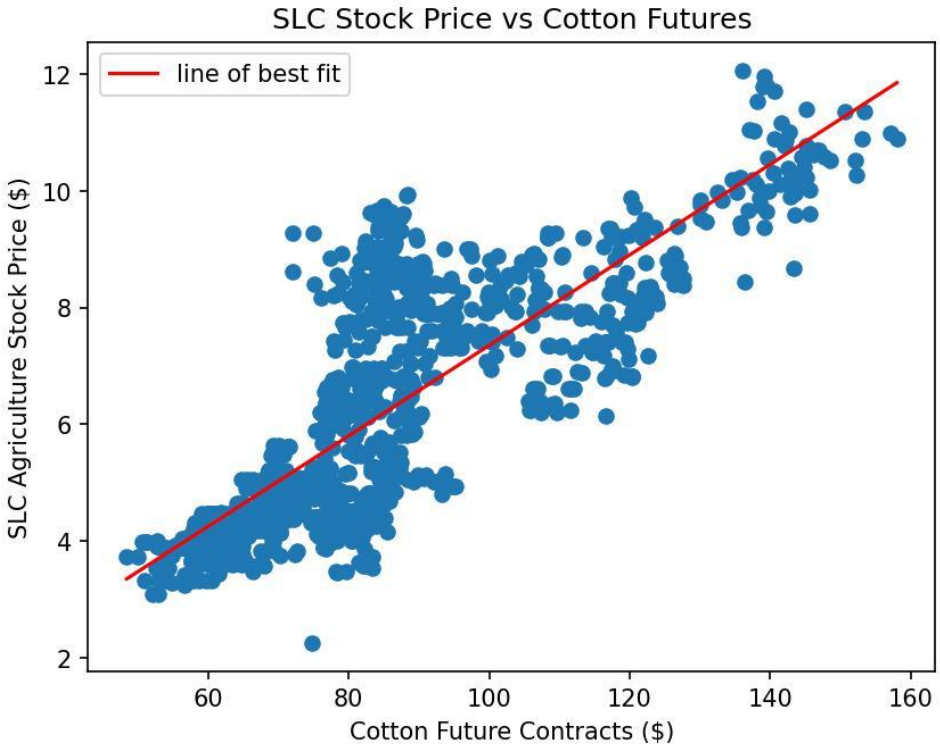
It is important to monitor the performance of the ML model and the overall pipeline in production. This can be done by tracking metrics such as accuracy and prediction latency, as well as logging any errors or warnings that occur.

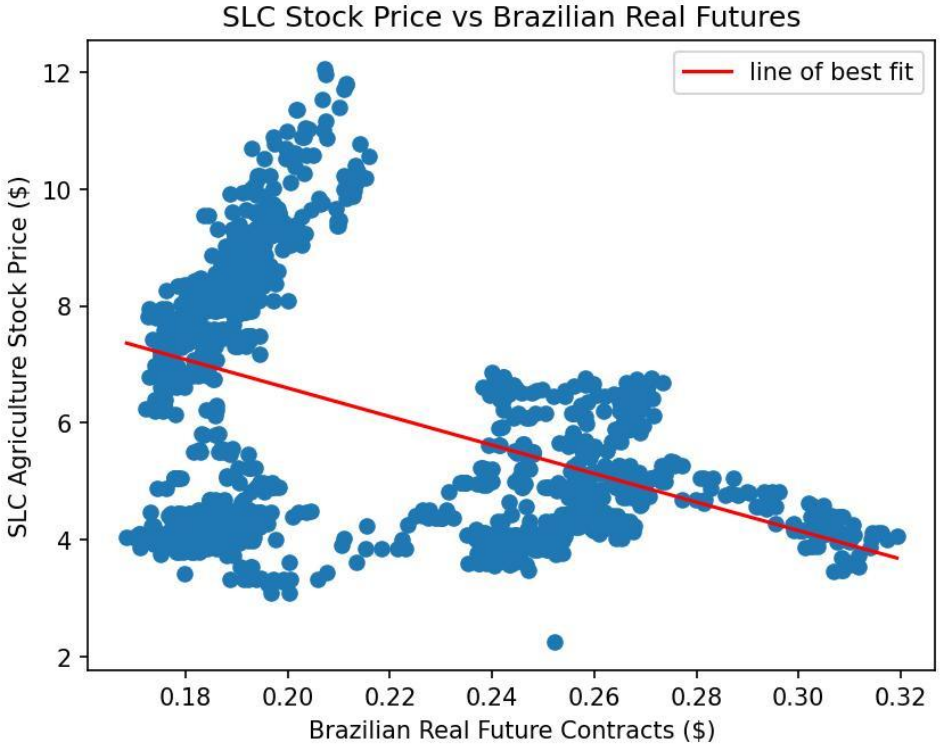
See **Appendix IV** for a diagram outlining the automation of the ML pipeline.

APPENDIX I: STOCK PRICE VS COMMODITY FUTURES & FX RATE CHARTS

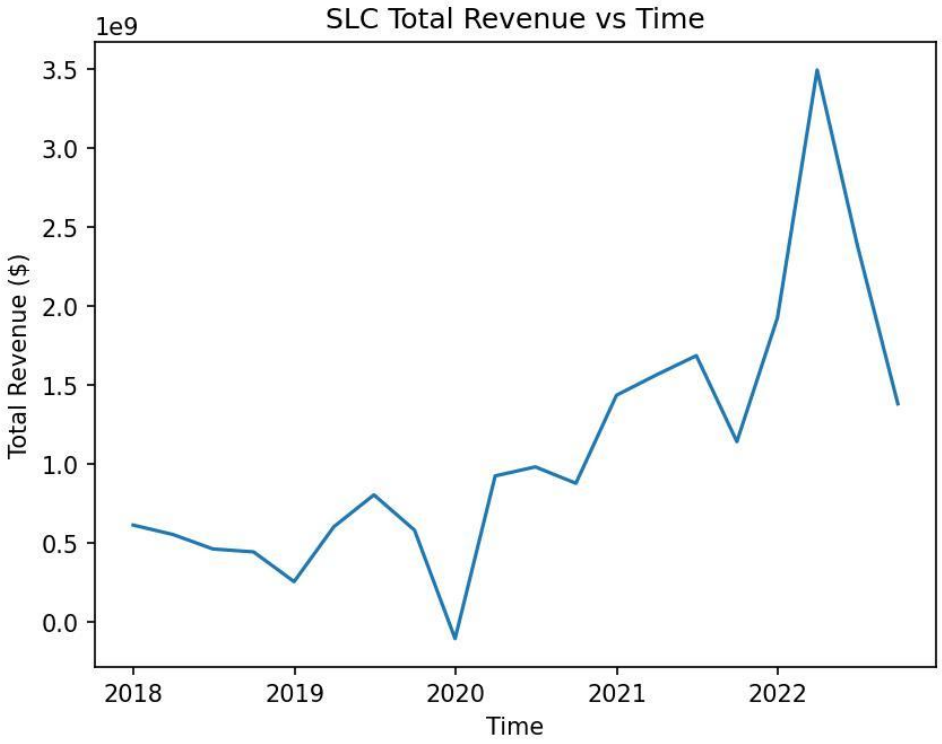

SLC Stock Price vs Soybean Futures
line of best fit
SLC Agriculture Stock Price ($)
Soybean Future Contracts ($)
2
4
6
8
10
12
800
1000
1200
1400
1600
1800

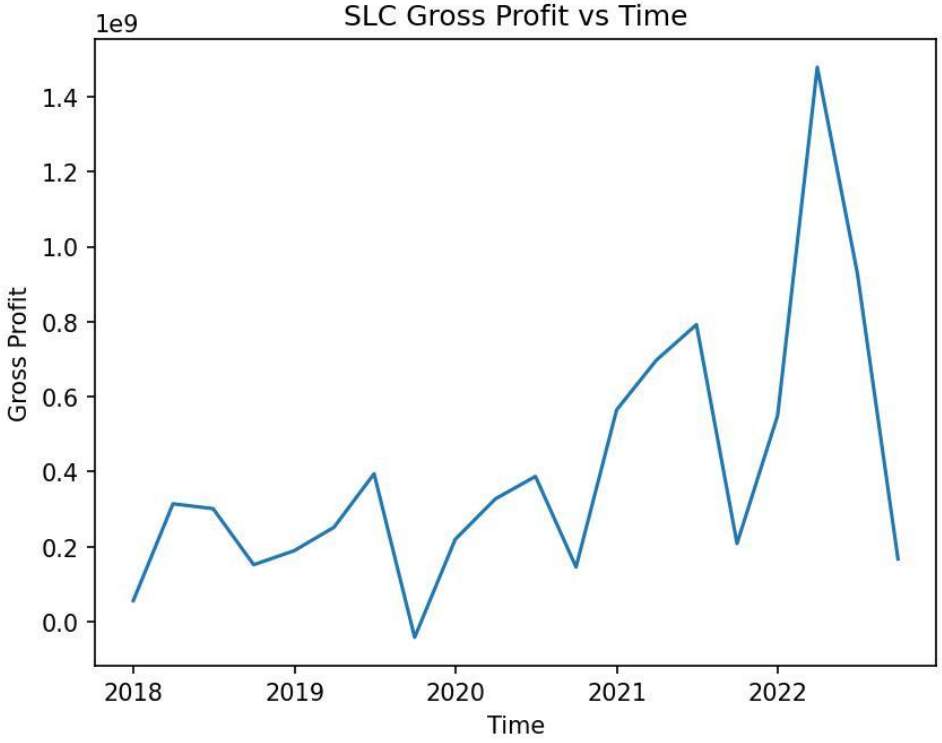

SLC Stock Price vs Corn Futures
line of best fit
SLC Agriculture Stock Price ($)
Corn Future Contracts ($)
2
4
6
8
10
12
300
400
500
600
700
800

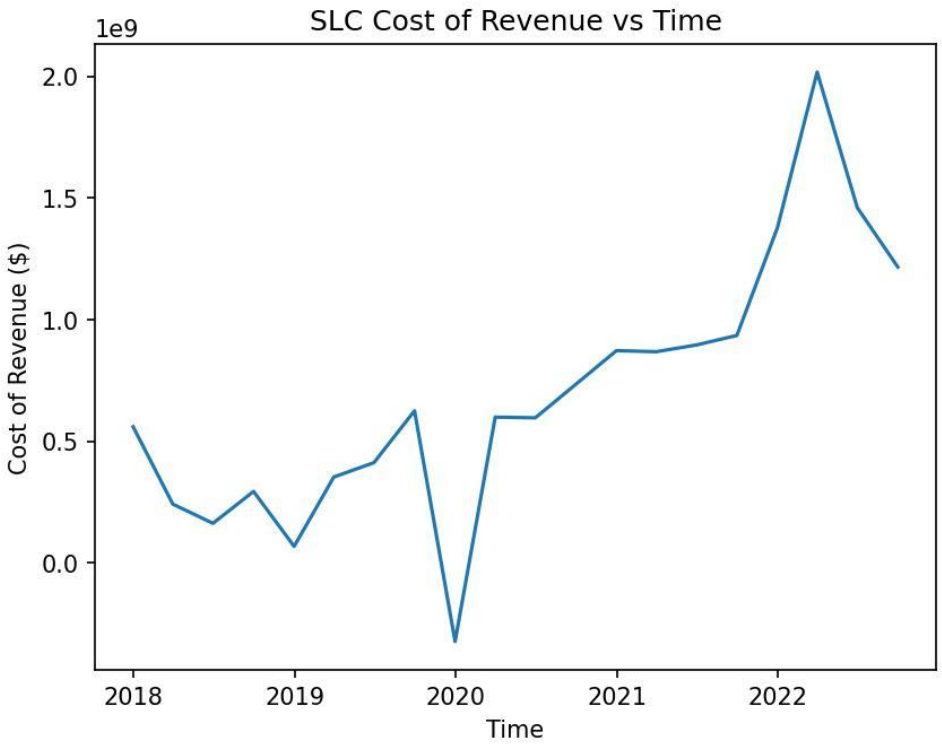

SLC Stock Price vs Cotton Futures
line of best fit
SLC Agriculture Stock Price ($)
Cotton Future Contracts ($)
2
4
6
8
10
12
60
80
100
120
140
160

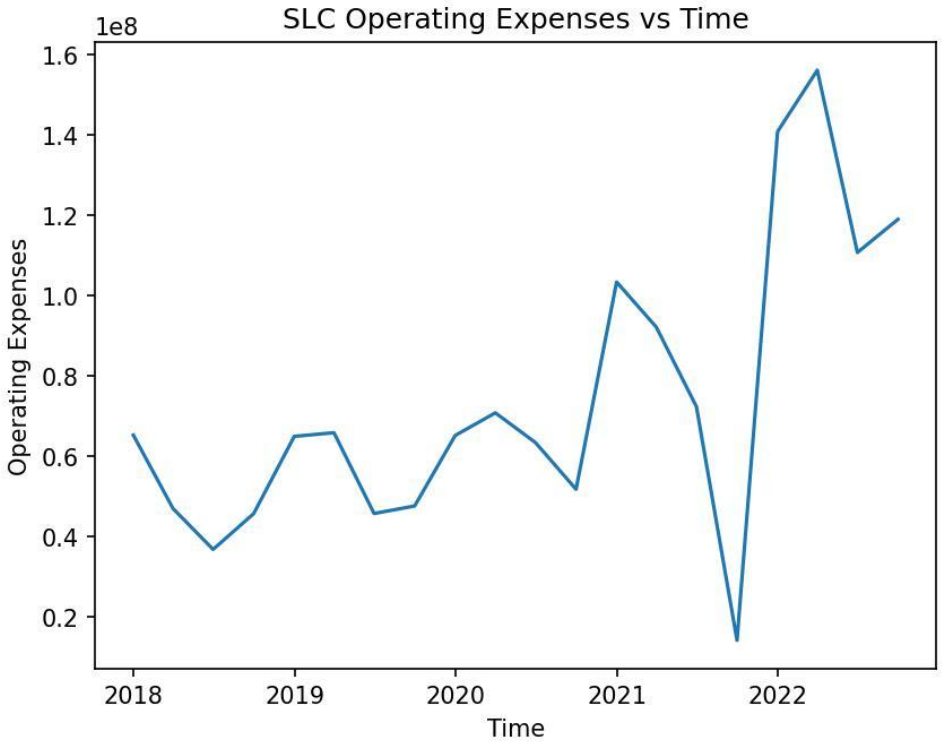

SLC Stock Price vs Brazilian Real Futures
line of best fit
SLC Agriculture Stock Price ($)
Brazilian Real Future Contracts ($)
2
4
6
8
10
12
0.18
0.20
0.22
0.24
0.26
0.28
0.30
0.32

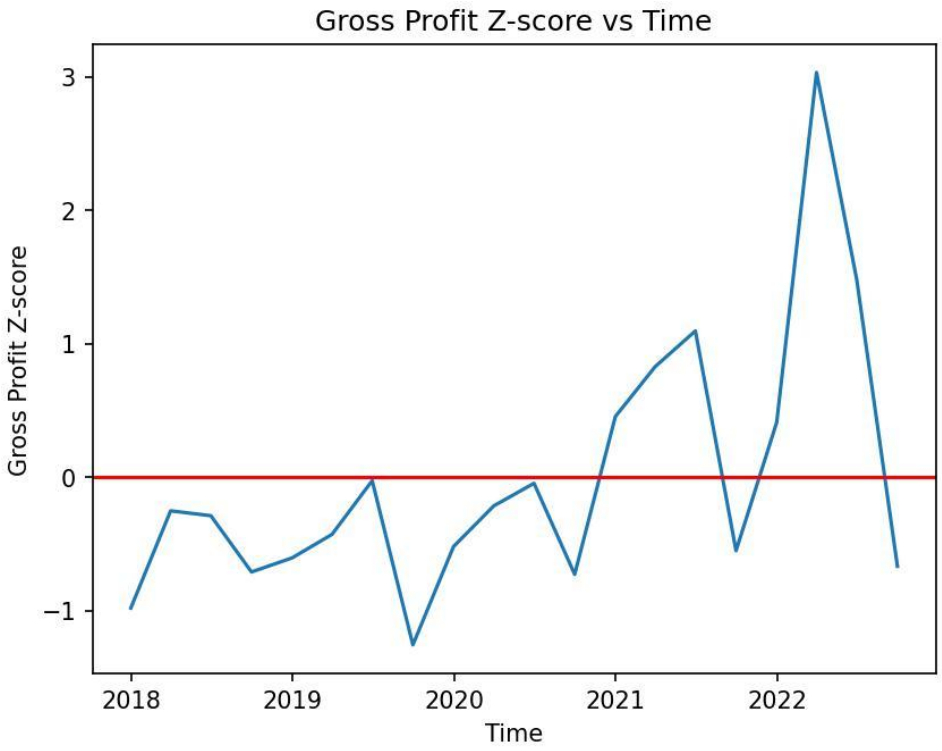
APPENDIX II: SLC AGRICOLA FINANCIAL PERFORMANCE CHARTS

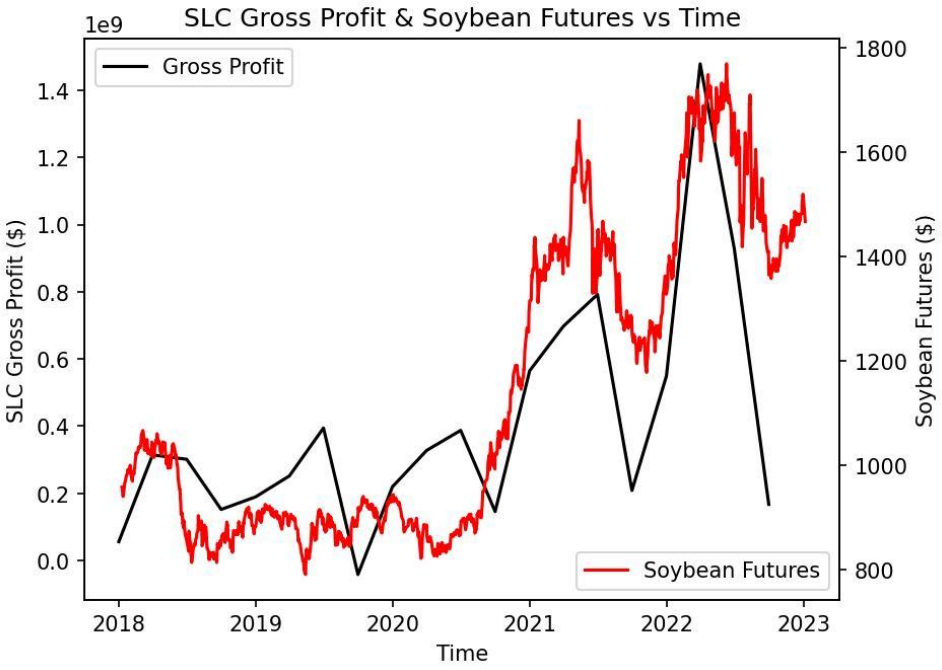

SLC Total Revenue vs Time
1e9
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Total Revenue ($)
2018
2019
2020
2021
2022
Time

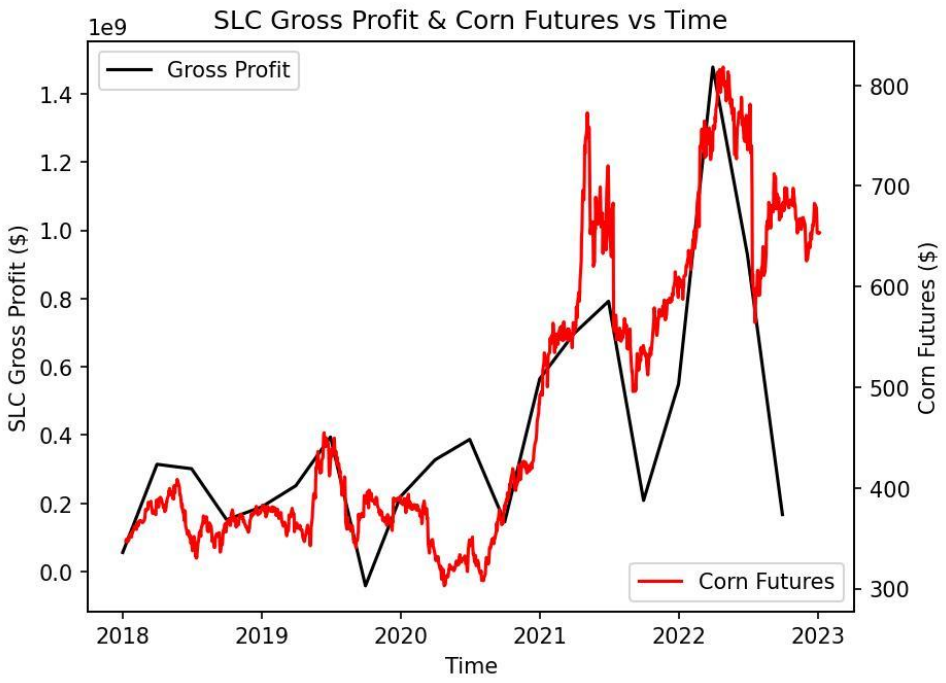

SLC Gross Profit vs Time
1e9
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Gross Profit
2018
2019
2020
2021
2022
Time

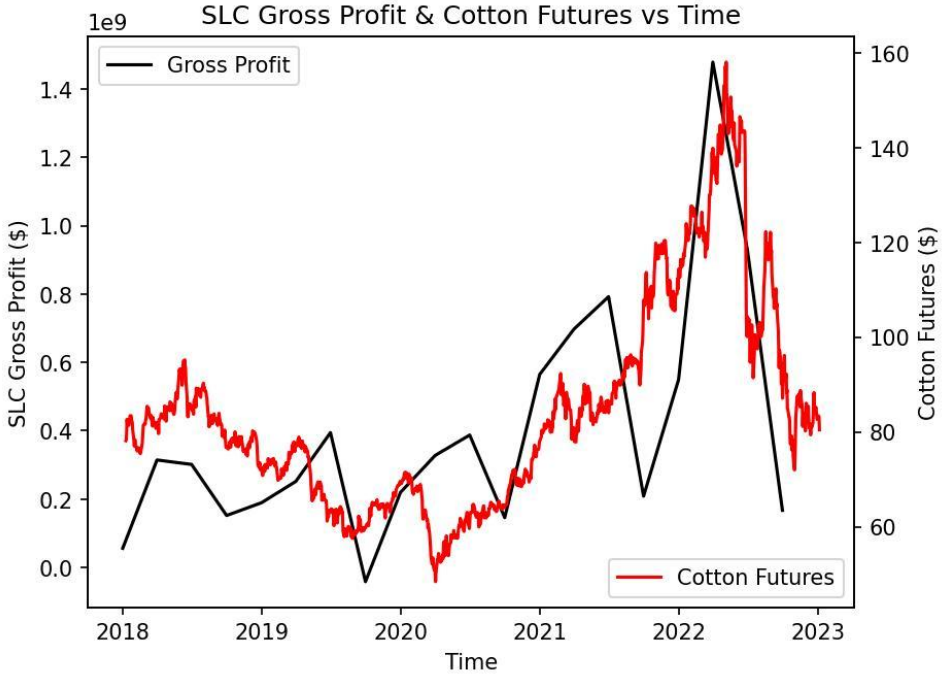

SLC Cost of Revenue vs Time
1e9
2.0
1.5
1.0
0.5
0.0
Cost of Revenue ($)
2018
2019
2020
2021
2022
Time


SLC Operating Expenses vs Time
1e8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Operating Expenses
2018
2019
2020
2021
2022
Time

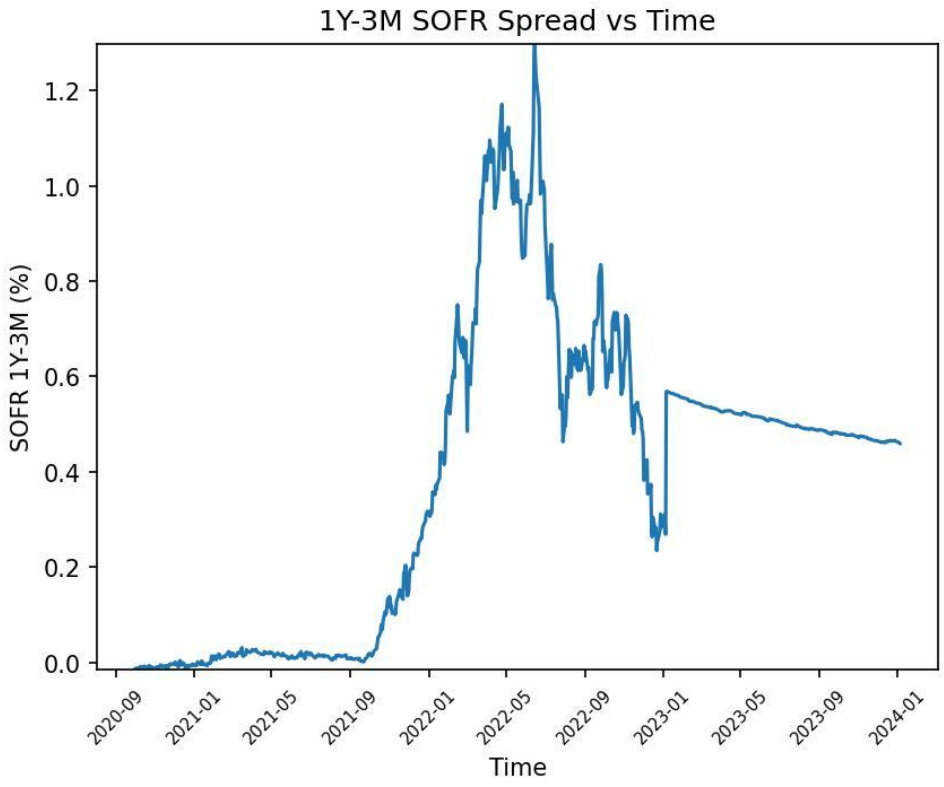

Gross Profit Z-score vs Time
Gross Profit Z-score
3
2
1
0
−1
2018
2019
2020
2021
2022
Time

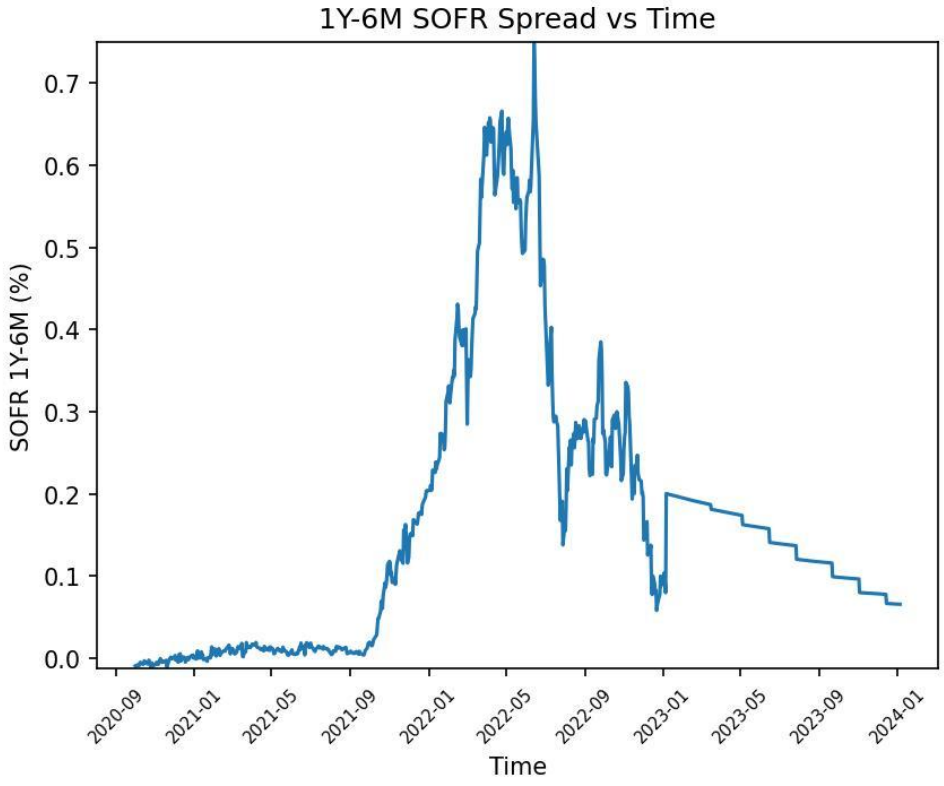

SLC Gross Profit & Soybean Futures vs Time
1e9
Gross Profit
Soybean Futures
SLC Gross Profit ($)
Soybean Futures ($)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1800
1600
1400
1200
1000
800
2018
2019
2020
2021
2022
2023
Time

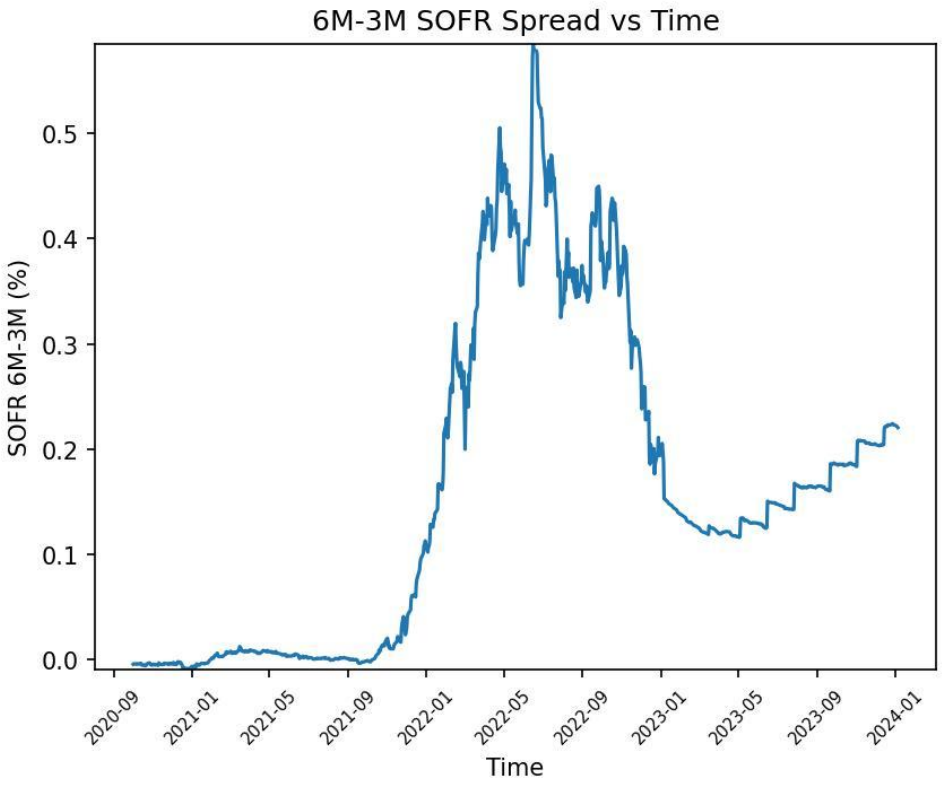

SLC Gross Profit & Corn Futures vs Time
1e9
Gross Profit
Corn Futures
SLC Gross Profit ($)
Corn Futures ($)
Time
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
300
400
500
600
700
800
2018
2019
2020
2021
2022
2023


SLC Gross Profit & Cotton Futures vs Time
1e9
Gross Profit
Cotton Futures
SLC Gross Profit ($)
Cotton Futures ($)
Time
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
60
80
100
120
140
160
2018
2019
2020
2021
2022
2023

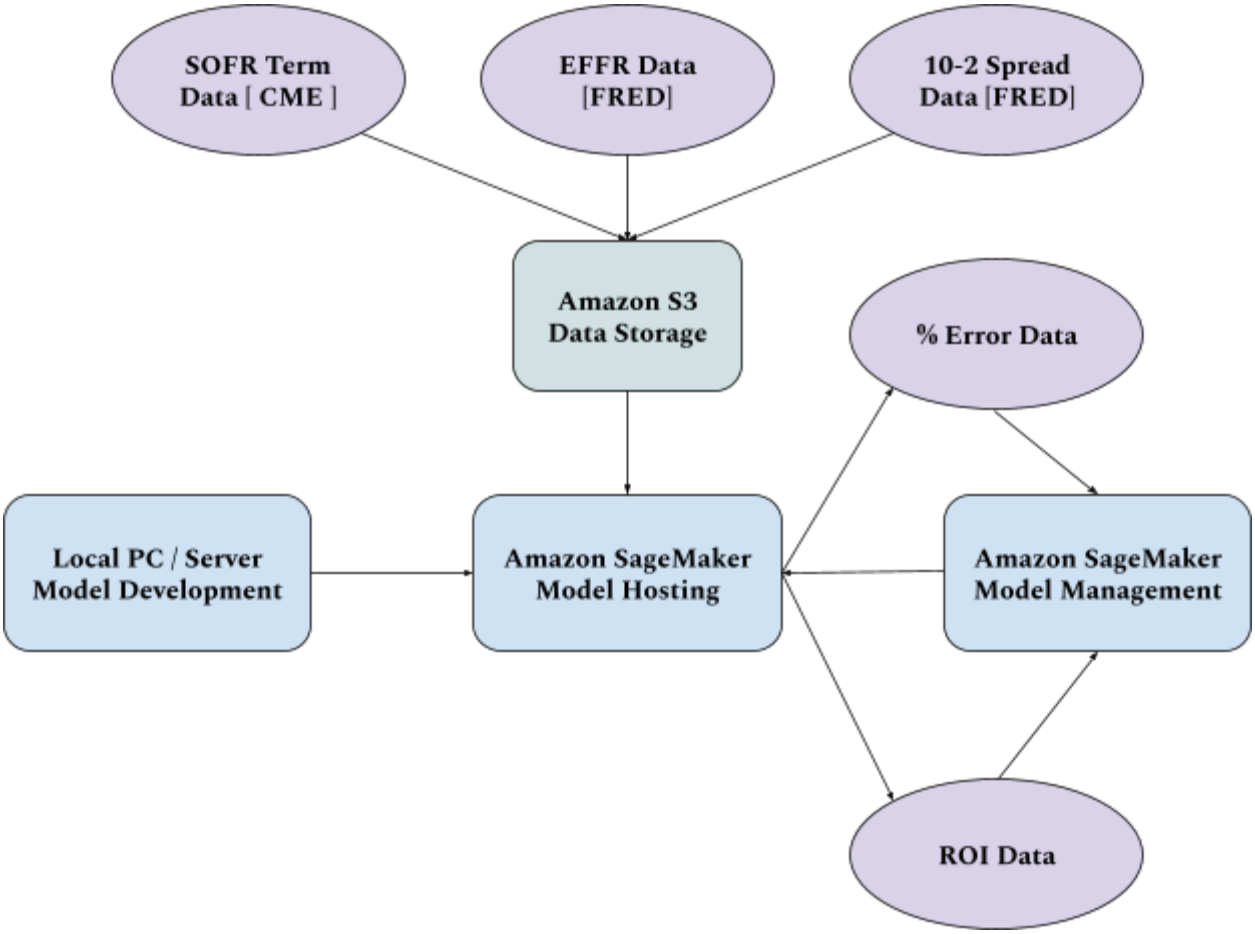
APPENDIX III: SOFR TERM SPREADS VS TIME CHARTS (INCLUDING PREDICTED VALUES)


1Y-3M SOFR Spread vs Time
SOFR 1Y-3M (%)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
2020-09
2021-01
2021-05
2021-09
2022-01
2022-05
2022-09
2023-01
2023-05
2023-09
2024-01
Time


1Y-6M SOFR Spread vs Time
SOFR 1Y-6M (%)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
2020-09
2021-01
2021-05
2021-09
2022-01
2022-05
2022-09
2023-01
2023-05
2023-09
2024-01
Time


6M-3M SOFR Spread vs Time
SOFR 6M-3M (%)
0.0
0.1
0.2
0.3
0.4
0.5
2020-09
2021-01
2021-05
2021-09
2022-01
2022-05
2022-09
2023-01
2023-05
2023-09
2024-01
Time

APPENDIX IV: ML PIPELINE AUTOMATION DIAGRAM


SOFR Term Data [CME]
EFFR Data [FRED]
10-2 Spread Data [FRED]
Amazon S3 Data Storage
% Error Data
Local PC / Server Model Development
Amazon SageMaker Model Hosting
Amazon SageMaker Model Management
ROI Data