Credit Spread Forecasting in Python

Giorgio Perego

April 2025

Contents

1 Introduction

The paper attempts to replicate and describe credit spread behavior with macroeconomic and market variables through the use of machine learning and market regime analysis in Python. It attempts to create a model that can forecast the direction of future credit spread based only on macroeconomic and market variables with high predictive power.

Credit risk and market risk are very closely related, as may be discerned from the observation that default probabilities and recovery rates track the business cycle. Furthermore, corporate bond spreads have been found to fluctuate with the expansion and contraction of the economy, in parallel with movements in aggregate and firm-level fundamentals. For example, seminal work by Fama and French (1989) illustrates that corporate bond yields increase as economic conditions weaken—reflecting higher credit risk. This report integrates both macroeconomic determinants and market-based proxies under a machine learning model. In doing so, we explain the so called "credit spread puzzle"—the persistent disconnect between theoretical estimates and empirical credit spreads—on the basis of more advanced modeling of default risk, interest-rate dynamics, and macro feautures.

In order to compare different approaches to modeling, we've set up two distinct data processing pipelines. The first one follows a more modern machine learning philosophy, with a focus on models such as Random Forest and XGBoost, and sparser transformations that facilitate quick adjustment to complex and nonlinear relationships in the data. The second pipeline, in contrast, follows a more academically traditional route—stationarity checks, factor analysis, and transformations—perfect for a Linear Regression framework, which we'll use as a baseline. Our analysis is conducted on daily credit spread values from the January 1st 2000 to the present.

All the data loading steps, transformations, and predictive modeling have been integrated into a single end-to-end Python pipeline to ensure reproducibility and real-world utility for future research on credit spread forecasting.

2 Background and Relevance of Credit Spread Analysis

Credit spread refers to the difference between the yield (return) of two debt securities with equal maturity but varying in credit rating. That is generally approximated as a corporate bond yield less the yield on a similar maturity government risk-free bond. The idea is that the credit spread is a measure of a security's risk premium, the additional yield an investor has to be paid to take on an additional credit risk. The idea is that investors have to be paid more (the spread) to make up for the added default risk of non-government bonds compared to government securities. The credit spread is computed as:

$$\begin{aligned} \textit{Credit Spread} &= \textit{Corporate Bond Yield} \\ &\quad - \textit{Treasury Bond Yield} \end{aligned}$$

Precisely in our code, it is computed by subtracting the yield of the 10-year U.S. Treasury bond (DGS10) from the BBB corporate bond yield (BAMLC0A4CBBBEY) — both of which are obtained from the Federal Reserve Economic Data (FRED).[1]

2.1 Relevance of Credit Spread Analysis

Being able to predict the credit spread can be useful as wider spreads indicate risk aversion and economic hardship, and narrower spreads indicate prosperity. In fact, we can monitor in way to foretells recessions or booms. Also, spreads measure the extra return required to hold risky bonds (i.e. corporates) in relation to government bonds. Widespread widening

[1] Using the BBB–10Year Treasury spread is a common and operationally efficient measure for ascertaining investor sentiment and risk pricing in the investment-grade corporate bond market.

shows fears regarding the capacity of corporations to refinance debt, particularly in situations of a slowdown or increase in rates. Furthermore, being able to predict the credit spread allows us to anticipate phases of expansion and recession. In fact, in an expansion phase, spreads narrow on the back of optimism on corporate earnings and high liquidity. Spreads widen during a contraction phase, predicting investment contraction, collapsing consumption, and GDP contraction. The 300 basis point increase from recent lows, for example, is a classic early warning signal.[2] In conclusion, credit spreads act as an early warning system, measuring investors' perception of systemic risk and the strength of the corporate sector, both of which are critical to economic stability.

2.2 Credit Spread High Yield and Investment Grade

We can distinguish between two types of credit spreads, associated with the different types of bonds to which they are referred. Indeed, we distinguish between credit spreads based on high-yield bonds (junk bonds or speculative grade) and investment grade bonds (safer with a rating higher than BBB). As regards the dissemination of high-yield credit, it encompasses those bonds rated below BBB- (or Baa3) and is linked to higher default risk. This is seen directly in the spread: on average, high-yield bonds have bigger yield spreads than investment grade bonds, in compensation for taking the higher risk. They mainly pick up the default risk and the credit quality of the issuer, largely driven by individual factors like corporate perception, liquidity, and recovery values in default. They pick up economic condition changes in an elevated way with comparatively higher sensitivity to the shape of the economic cycle and perceived risk. In times of financial stress, they grow a lot, indicating the increased fear of default. On the other hand, when there are economic upticks and universal access to credit, they deflate considerably.

Conversely, investment grade credit spreads, which are comprised of securities rated from BBB to AAA (or their equivalents), thus carrying a low or intermediate probability of default, are macro-driven rather than idiosyncratic risks. Usually more exposed to macroeconomic environment, monetary policy, interest rate direction and corporate fundamental stability, they are more interest rate sensitive to movement, more sensitive to movement of risk-free rate (e.g. Fed cuts) and yield curve. They are less sensitive and more stable in normal market conditions, since investment grade issuers are considered to be more stable. But even in times of liquidity crisis or recession, spreads can rise but in a comparatively moderate way against the high yield sector.

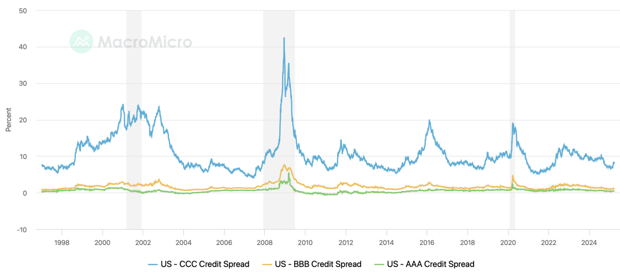

Figure 1: *Credit spreads for CCC, BBB, and AAA rated U.S. corporate bonds from 1996 to 2024 (percent)*[4]

We employed in our research the spread between US Treasuries and BBB rated (or Baa rated in Moody's rating) bonds because they are the lowest investment grade level, and thus a good and utilized proxy for the market. Also, this means that the relative spread is highly sensitive to changes in risk perception on the market. BBB bonds provide a representative picture of the bond market. In fact, the majority of medium-high capitalization companies actually belong to this category, and thus the

[2]In 2007, widening spreads preceded the subprime mortgage crisis. In 2020, soaring spreads anticipated the COVID-19 recession.

[4]The shaded areas indicate recession periods, highlighting how high-yield (CCC) spreads are notably more volatile than investment-grade (BBB, AAA) spreads.

BBB spread is a good "synthesis" indicator of financing conditions for companies. Finally, compared to higher rating spread (e.g. AAA), the BBB is more responsive to macroeconomic and market shocks but less than high yield, hence making a perfect proxy of the bond market.

3 Data Description and Preprocessing

3.1 Data Description

As would be anticipated before analysis, we have the credit spread as the 10-year U.S. Treasury bond yield (DGS10) minus the BBB corporate bond (BAMLC0A4CBBBEY). In this present study we construct an aggregated dataset from multiple data sources to include macroeconomic as well as market variables that are relevant to credit spread analysis. In particular, we use four main sources of data to create a solid dataset for our credit spread analysis. The first is *Federal Reserve Economic Data* (FRED), which is one of the leading providers of economic and financial data, where we extracted the *BBB corporate bond yield* (ICE BofA BBB US Corporate Index - BAMLC0A4CBBBEY) and the *10-year Treasury yield* (DGS10). Out of their difference we obtain our target variable.

In addition to the bond yield data, we also include a group of auxiliary variables designed to capture more pervasive economic and market conditions. As an example, we use the *3-Month Treasury Bill Secondary Market Rate* (TB3MS) useful in approximating the short-term money cost and market expectations regarding monetary policy. Actually, it is even able to influence the credit spread and the bond yields; short-term rate increases are actually able to indicate monetary tightening and increase the perceived risk. The *Terms of Credit, Auto Loans, 48-Month New Car* (TERMCBAUTO48NS) is also beneficial to our purpose because it has been shown to be a good proxy for the real sector credit conditions, which is to be related to corporate credit risk. It is actually able to measure the cost of consumer credit and serve as an indicator of demand and economic prosperity.

The *Personal Consumption Expenditure* (PCE) and *Consumer Price Index for All Urban Consumers* (CPIAUCSL) are then used as indicators of inflation. The first is FRED's default measure of the trend of inflation, which over the course of time worsens the real value of bond yields and impacts the credit spread, and the second, favourite among the media and investors, provides another reading of inflation useful to compare with PCE and gauge market expectations.

Finally, for a general analysis, the *GDP* is also considered, which can quantify economic growth; a low GDP can mean recession and increase the perceived risk.

As second data source we use *Yahoo Finance* from which we extract data for the *S&P 500 ETF* (SPY) and derive a momentum proxy by computing a difference between two exponential moving averages (EMA12 and EMA26). We get a time series, *SPY_DIFF* (EMA12 - EMA26), that is a market sentiment leading indicator, which shifts each day. These two series can be very helpful as technical indicators of momentum in the stock market. EMA12 is short-term momentum (about 2 weeks of trading) with a quick reaction to new momentum in the market and EMA26 is a medium trend (about 1 month of trading), the slower, less volatile view. With the use of both, the momentum can be determined. We also include these variables, SPY and SPY_DIFF, within our model due to the reason that market mood also influences credit spread. A $SPY\_DIFF > 0$ represents a bull market when the S&P 500 increases, the mood is optimistic, perceived risk decreases, and the credit spread will compress and a $SPY\_DIFF < 0$ represents a bear market when the S&P 500 decreases, uncertainty increases, and the credit spread widens.

In addition, our analysis considers the *TED*

Spread, straight from *FactSet*, i.e. the 3-month LIBOR rate minus the 3-month Treasury Bill rate. It is a measure of interbank market stress and its high value indicates financial tension (e.g. 2008 crisis), which has a broadening effect on the credit spread.
Finally, our project also employed macro political uncertainty indicators that reflect the overall stability of US policy such as Financial Regulation and Fiscal Policy (*Categorical Economic Policy Uncertainty Data*).[5]

All information ranges from January 1, 2000 through to the most current available date to provide an adequate sample for in-sample training as well as out-of-sample testing. The final merged DataFrame used as input to our two model pipelines, resulting in a robust dataset blending macroeconomic factors, market sentiment proxies, policy indicators, and credit yield information to provide an end-to-end overview of the potential drivers of credit spread evolution over time.

3.2 Data Trasformation

The combined dataset is quite unclean with several issues including missing values, varying scales, non-stationarity and non-normal distributions. To give our model better quality data, we followed the below process to preprocess and clean the dataset. In particular, we distinguish between two different pipelines, which address different data manipulation as they work with models with different requirements. In fact, pipeline 1 will be utilized as a dataset to train non-linear models such as Random Forest and XGBoost while pipeline 2 will be utilized as a starting point for a Linear Regression, a more traditional model that will be utilized as a benchmark to verify how newer models are more efficient.

3.2.1 Missin Value

There are several reasons for missing values in our data. Because of that we address them during the loading by removing rows that lack essential information. Furthermore, during both the merging of the dataset and the synchronization of the lower frequency series with the daily ones using forward fill methods, that preserve the last known value until new data arrives, we remove any remaining observation with missing values.

3.2.2 Market Regime

The "market regime" is the different states or phases the market can be in over a time horizon. As an example, a market regime can be associated with periods of high volatility, periods of low volatility, rising markets, falling markets or some other special forms of behaviour. By using a model such as the GaussianHMM (Hidden Markov Model), you are able to segment the market whereby the model can automatically differentiate and segment the market behaviour into different regimes.

Every regime will have a different state, which may have its own features (such as wide spreads during periods of financial turmoil or tight spreads during periods of calm).
The model can now enhance the prediction model by introducing the "market regime" as a feature as the contextual model provides added substance. The model being able to know which regime the market is in, makes the predictive model place into proper perspective the fluctuations of the target. [6]

[5] See Table 3 for complete data

[6] For both pipelines, that are aimed for two different destinations, the regime of the market is calculated similarly but not the same. The first one, that is pointed towards a larger exploratory and predictive analysis, uses the regime of the market as a tool to segment data and analyze how the Credit Spread behaves among different states in the market. The second pipeline is more committed to a linear simplified forecast with a systematic framework that includes factor analysis and stationarization. The market regime is included as a special characteristic to enhance a linear regression model.

Table 1: Regime Analysis

Regime	Mean	Std	Count
0	1.534273	0.110000	571
1	5.296320	1.468722	250
2	1.916177	0.097387	769
3	1.533063	0.107797	555
4	1.117887	0.202696	1898
5	2.476161	0.248573	995

Table 1 shows that the sample period adopted (2000-2025) was predominantly characterized by normal times. In fact, Regime 4 with low spread and low volatility has the largest number of observations. Regime 2 with high spread and high volatility is less frequent but accounts for extreme occurrences such as the 2007-2008 and 2020 crises (see Figure 2) while Regime 0,2,3,5 display intermediate regimes between the two.

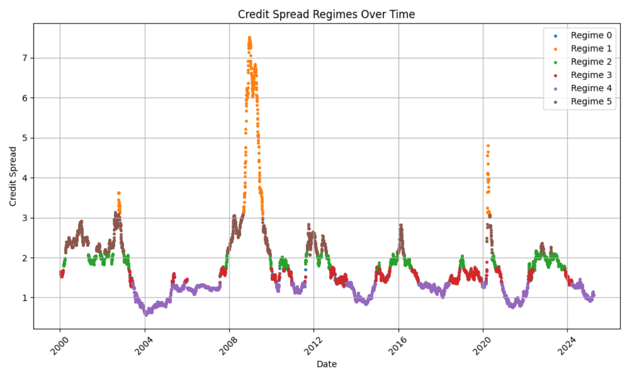

Figure 2: Credit Spread Regime

3.2.3 Skewness reduction and Standardization

In our implementation, we address skewness and normalization of features in both pipelines to have an appropriate data ready for the machine learning models used. The step is necessary to minimize the impact of the asymmetric distributions and we normalize to make the features at the same scale. In particular, each of the variables is checked and if their skew is greater than the threshold of 0.75, they are transformed. Finally, since our features differ in size, they are standardized and normalized using the formula:

$$z = \frac{x - \mu}{\sigma}$$

where x is the original value, μ is the mean, and σ is the standard deviation. [7]

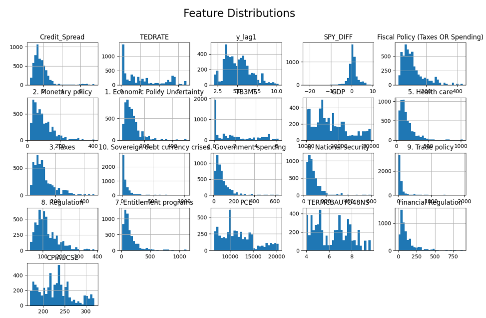

Figure 3: Feature distribution

3.2.4 Factor Analysis

For machine learning algorithms, in themselves, multicollinearity and redundancy would not be an issue in prediction but may cause incorrect parameter estimation. Due to this, in Pipeline 2, in order to manage this problem, a factor analysis is used, which is a statistical technique used to reduce the dimension of a set of observed variables to fewer latent factors that explain the common variance among the variables. This process is thus able to reduce the dimensionality and deal with the multicollinearity demonstrated by the heatmap.

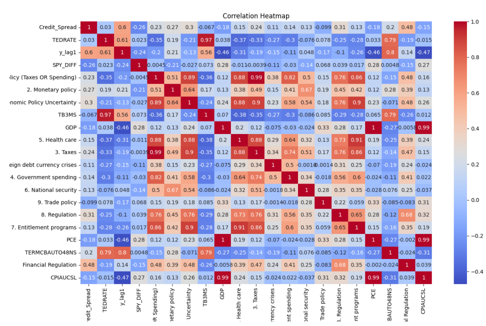

Figure 4: Correlation Heatmap

[7] Figure 3 in particular shows the distribution of the various features showing how they have a tail on the right, a characteristic sign of a right-skewed distribution.

[8] Then, the stationarity of the features is established using the Augmented Dickey-Fuller (ADF) test, then applying differentiations until the series become stationary (p-value $<$ 0,05). Finally, a Random Forest is used to select features with importance $>$ 0,005. If two features are redundant, then likely one of them is not as important and is removed.[9]

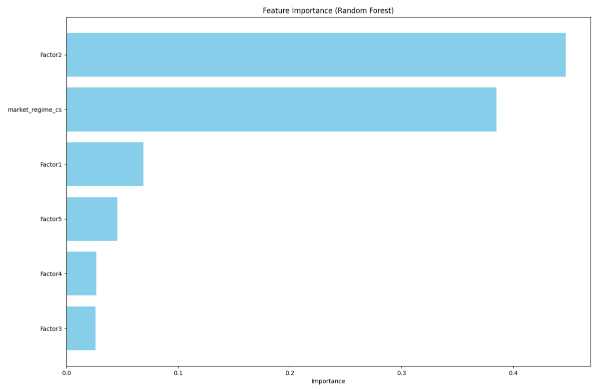

Figure 5: Factor Importance

4 Model

For all the model below, we split the dataset into two subsets with a ratio of 80/20 between training and testing and then we use the Root Mean Squared Error (RMSE) and the Coefficient of Determination (R^2) as performance metrics. Also for the non-linear models we tune the parameters.

4.1 Linear Regression

The first model we use is Linear Regression. This model would serve as benchmark to evaluate another models' performance. Linear Regression is a parametric model that assumes a linear relationship between features and the target. It relies on the least squares method to estimate the coefficients that best describe this relationship. Specifically in linear regression a response variable Y is modelled as a linear function of k predictor or explanatory variables, following the formula:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon.$$

The linear regression model may be written in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

But this model and its precision are the result of heavy operations on the dataset, which violate the data's reality. Actually, it exploits the linear structure of the pre-treated data (stationary variables), giving simplicity and interpretability. Also, financial markets are famous with its non-linearity. We believe a linear simple model may fail to model some deep patterns of the financial data. We would then employ tree-based models to further tune our prediction.

4.2 Random Forest

Random Forest is an ensemble algorithm based on decision trees that increases robustness and predictability by aggregating the predictions of numerous trees.

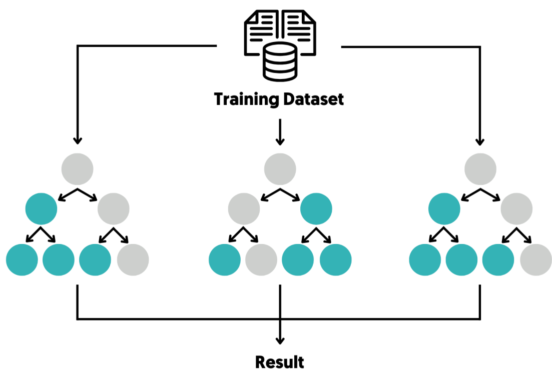

Figure 6: Random Forest simplified

A random subset of the features is utilized at each split in the creation of each tree. This method, apart from reducing the correlation between trees, increases the model's ability to identify non-linear relationships and interactions.

[8]Factor1: "Policy and Fiscal Uncertainty", Factor2: "Credit Spread Factor", Factor3: "Sovereign/Global Risk Factor", Factor4: "Interest Rate Factor", Factor5: "Trade Policy Factor"

[9]Linear regression assumes that the data is stationary. If the features have trends or variable variance, the model may produce false or unstable results.

4.3 eXtreme Gradient Boosting

XGBoost, or Extreme Gradient Boosting, is a boosting procedure in which the model is built sequentially with the new tree trying to correct the residual errors of the previous trees. The method is based on gradient descent-based loss function optimization, progressively enhancing the prediction power of the model.

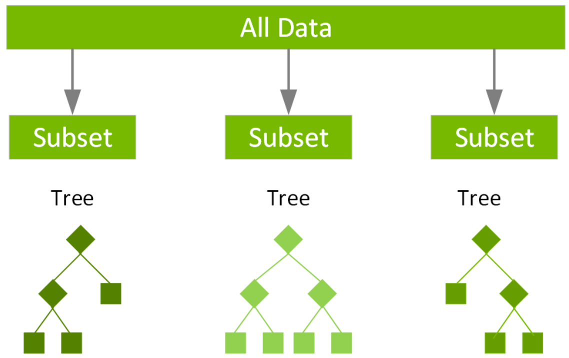

Figure 7: XGBooost simplified

5 Results

The results show that Random Forest and XGBoost are on par with each other, which proves that financial data are not linear in nature. Random Forest is slightly better than XGBoost by a very small margin— mostly allegedly due to its less intricate structure and noise resilience— but an excellent MAE shows that the model is also not influenced by outliers. XGBoost, although very powerful in modeling complex patterns by sequential boosting, might have been penalized by a still suboptimal tuning (very expensive at computational level).

Therefore, the similarity between Random Forest and XGBoost lies in the fact that both models are decision tree-based ensemble models with the capacity to learn complex interactions and non-linear relationships, but Random Forest has a simpler structure and less hyperparameter sensitivity, while XGBoost requires more attention to tuning to avoid overfitting or underfitting. [10]

Linear Regression is the worst performing model, highlighting why, despite stringent preprocessing step, it may never be able to capture the undercurrent complexity in market behavior. Yet, considering its simplicity and operating on the transformed scale, it nonetheless exhibits a commendable level of performance.

Model	RMSE	MAE	R^2
Random Forest	0.048047	0.035713	0.989208
XGBoost	0.048549	0.038291	0.988982
Linear Regression	0.052959	0.049316	0.927080

Table 2: Performance metrics for the evaluated models

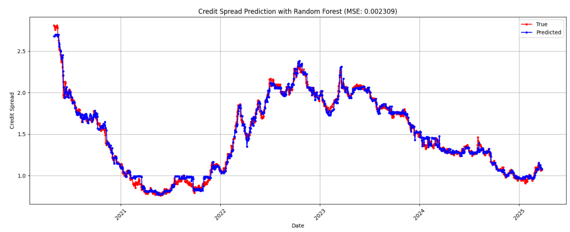

Figure 8: Random Forest Results

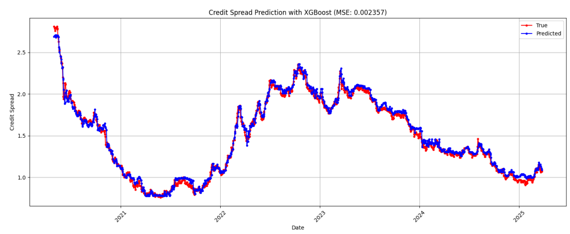

Figure 9: XGBoost Results

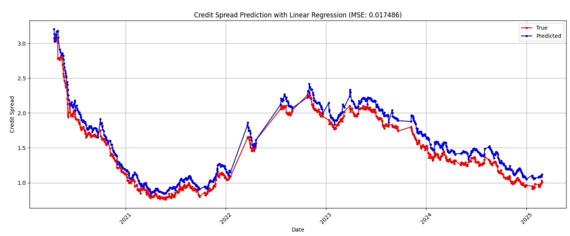

Figure 10: Linear Regression Results

[10] Our model tuning process optimizes both Random Forest and XGBoost for predicting credit spreads. For Random Forest, we use RandomizedSearchCV with TimeSeriesSplit to explore hyperparameters employing EarlyStoppingRandomForest to select the optimal number of trees based on a validation set. For XGBoost, BayesSearchCV tunes parameters and subsample over multiple iterations.

6 Conclusion

In this project, we have predicted credit spread based on macro-level, bond-level, and sentiment features combined. We have decided to use two pipelines so that the models can be run in the most professional and efficient way. Empirical results showed that Random Forest and XGBoost as machine learning-based models outperformed a baseline linear regression model. The linear model, despite being used as a baseline, proved to be severely lacking in terms of being able to capture the underlying market complexity, even if favored by an optimal preprocessing step.

Our study specifies the advantage of ensemble models for predicting credit spreads and shows that, in cases with noisy data and non-linear relationships, approaches such as Random Forest can generate more stable and accurate estimates.

For future implementations it would be interesting to exploit this high predictive capacity to develop trading strategies such as, for example, if the model predicted a widening of credit spread, a short high yield or corporate ETFs and long Treasuries trade can be hypothesized, or corporate bonds can be overweighted relative to government bonds.

References

[1] Fama, E. F., & French, K. R. (1989). Business Conditions and Expected Returns on Stocks and Bonds. *Journal of Financial Economics*

[2] Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*

[3] Tang, D. Y., & Yan, H. (2010). Market conditions, default risk and credit spreads. *Journal of Banking & Finance*

[4] Federal Reserve Bank of St. Louis. (n.d.). *FRED Economic Data*. Retrieved from `https://fred.stlouisfed.org/`

[5] Yahoo Finance. (n.d.). *Yahoo Finance*. Retrieved from `https://finance.yahoo.com/`

[6] FactSet Research Systems Inc. *FactSet Data Services.*