Unlocking Pre-Provision Net Revenue: Diving into the Power of Interest Income

Mentor: Dr. Kristina Martin
Harshita Agrawal, Meaghan Allen, Obed Domson, Yixuan Huang, Sebastian Jaramillo

July 26, 2024

Executive Summary

In the aftermath of the 2008 financial crisis, the Federal Reserve established stress tests to evaluate banks' resilience against economic downturns. A critical component of these stress tests is the assessment of Pre-provision Net Revenue (PPNR), which measures a bank's financial performance by considering interest income, noninterest income, interest expenses, and noninterest expenses. This project aims to enhance the modeling of PPNR, particularly focusing on interest income, by leveraging advanced machine learning techniques.

Our research utilized publicly available financial data from the Federal Reserve, specifically examining quarterly reports from banks. We transformed this data to derive key metrics such as quarterly interest income, which was then normalized for accuracy. We also incorporated various macroeconomic indicators, including the yield curve, GDP, and unemployment rate, to understand their impact on interest income.

We explored several modeling approaches to predict interest income. Initially, linear regression models were used, revealing that incorporating bank-specific factors significantly improved prediction accuracy, with an R^2 score of 0.79. However, adding macroeconomic indicators only marginally enhanced performance. Next, regression trees were applied, which effectively handled non-linear relationships and confirmed the importance of lagged income as a predictor. Autoregressive models provided insights into how past values influence future predictions, identifying the optimal model order to minimize errors. Finally, random forests, which aggregate multiple decision trees, demonstrated the highest accuracy with an R^2 score of 0.81, proving robust in handling complex datasets and relationships.

The findings indicate that integrating macroeconomic indicators with bank-specific characteristics produces the most accurate predictions for interest income. The random forest model, in particular, offered superior performance and reliability. This research not only validated the effectiveness of these models but also highlighted the potential for future improvements. Moving forward, extending the dataset or exploring additional components of PPNR could further enhance the precision of financial performance models and stress-testing methodologies.

Keywords— Pre-provision Net Revenue (PPNR), Machine Learning Techniques, Interest Income Modeling, Random Forests, Macroeconomic Indicators

1 Introduction

The Federal Reserve System is the governing bank in the United States. Their five key functions are to conduct the nation's monetary policy, to promote the stability of the financial system, to promote the safety and soundness of individual financial institutions, to foster payment and settlement system safety and efficiency, and to promote consumer protection and community development [3]. After the 2008 financial crisis, the Federal Reserve started implementing a modern stress test. The goal of the stress test is to assesses banks' ability to withstand adverse economic conditions and to absorb losses during stressful conditions [2]. This is useful because it gives security to people that even if a large financial crisis happened again, the banks will still be alright. Since the 2008 crisis and with the help of stress testing, banks were able to successfully withstand the impacts of the Covid-19 Pandemic in 2020.

One important aspect of stress testing is modeling Pre-provision Net Revenue (PPNR), which is a measure of a bank's performance. The PPNR of a bank is calculated with the following formula:

$$\text{PPNR} = \text{Interest Income} - \text{Interest Expense} + \text{Noninterest Income - Noninterest Expense.}$$

To project the various components of PPNR, the Federal Reserve uses different types of models such as autoregressive models, simple nonparametric models, and structural models [3].

In our project, we will be focusing on one component of PPNR, namely Interest Income. This was chosen due to the explainability of macroeconomic variables that affect it. For example, it makes sense that yield curves will affect interest income because it affects how much interest a bank will charge for an interest-bearing asset, therefore affecting how much profit they will make. We will be looking to expand modeling of PPNR components by making new models for interest income with machine learning.

2 Description of Data and Data Visualization

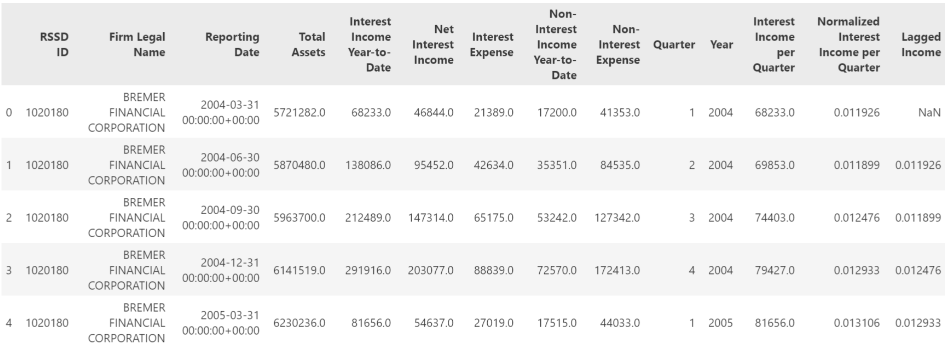
2.1 Financial Data

We used public data from the Federal Reserve FR-9YC form. The FR-9YC form collects basic financial data from U.S. bank holding companies (BHC), savings and loan holding companies (SLHC), U.S. intermediate holding companies (IHC) and securities holding companies (SHC) that satisfy a threshold for total assets (currently this threshold is at $ 3 billion) [1]. This form is collected quarterly and contains information about each bank, their income and expenses, their assets, their loans, and more.

In this project, we downloaded this data from 2004 up to 2024 quarter 2. This was a large dataset (over 2000 variables), so we first took a subset of this data that only included the variables we were interested in – namely bank information data, reporting dates, total assets, and the components of PPNR. The variable we were interested in, interest income, was reported as year-to-date, while the other variables in the dataset were reported as quarterly. To overcome this difficulty, we made the following code to create a new variable which indicated the interest income per quarter.

```
Function f(val):
    global last_val
    new_val = val - last_val
    last_val = val
    return new_val

Initialize df as an empty list

For each unique RSSD_ID in dataframe:
    Filter dataframe by current RSSD_ID into df_i

    For each unique Year in df_i:
        Filter df_i by current Year into df_i_j
        Sort df_i_j by Quarter
        Set last_val to 0
        Apply function f to "Interest Income Year-to-Date" column in df_i_j
        Store results in "Interest Income per Quarter" column
        Append df_i_j to df

Combine all dataframes in df into new_dataframe
```

Next, we needed to normalize the interest income per quarter by dividing our new variable interest income per quarter by total assets. Once that was complete, we made another new variable called lagged income, which gave the previous quarters normalized interest income per quarter. The picture below shows a small subset of our completed dataset.



	RSSD ID	Firm Legal Name	Reporting Date	Total Assets	Interest Income Year-to-Date	Net Interest Income	Interest Expense	Non-Interest Income Year-to-Date	Non-Interest Expense	Quarter	Year	Interest Income per Quarter	Normalized Interest Income per Quarter	Lagged Income
0	1020180	BREMER FINANCIAL CORPORATION	2004-03-31 00:00:00+00:00	5721282.0	68233.0	46844.0	21389.0	17200.0	41353.0	1	2004	68233.0	0.011926	NaN
1	1020180	BREMER FINANCIAL CORPORATION	2004-06-30 00:00:00+00:00	5870480.0	138086.0	95452.0	42634.0	35351.0	84535.0	2	2004	69853.0	0.011899	0.011926
2	1020180	BREMER FINANCIAL CORPORATION	2004-09-30 00:00:00+00:00	5963700.0	212489.0	147314.0	65175.0	53242.0	127342.0	3	2004	74403.0	0.012476	0.011899
3	1020180	BREMER FINANCIAL CORPORATION	2004-12-31 00:00:00+00:00	6141519.0	291916.0	203077.0	88839.0	72570.0	172413.0	4	2004	79427.0	0.012933	0.012476
4	1020180	BREMER FINANCIAL CORPORATION	2005-03-31 00:00:00+00:00	6230236.0	81656.0	54637.0	27019.0	17515.0	44033.0	1	2005	81656.0	0.013106	0.012933

2.2 Macroeconomic Data

Interest income fluctuates in response to changes in economic conditions over the nine quarters of the projection horizon. These conditions are influenced by various macroeconomic variables. In our analysis, we have considered several key variables, including, Yield Curve and Yield Spread, Gross Domestic Product (GDP) and GDP Growth, and Unemployment Rate. These variables collectively impact interest income, and their interactions are integral to understanding and projecting economic outcomes over the specified period.

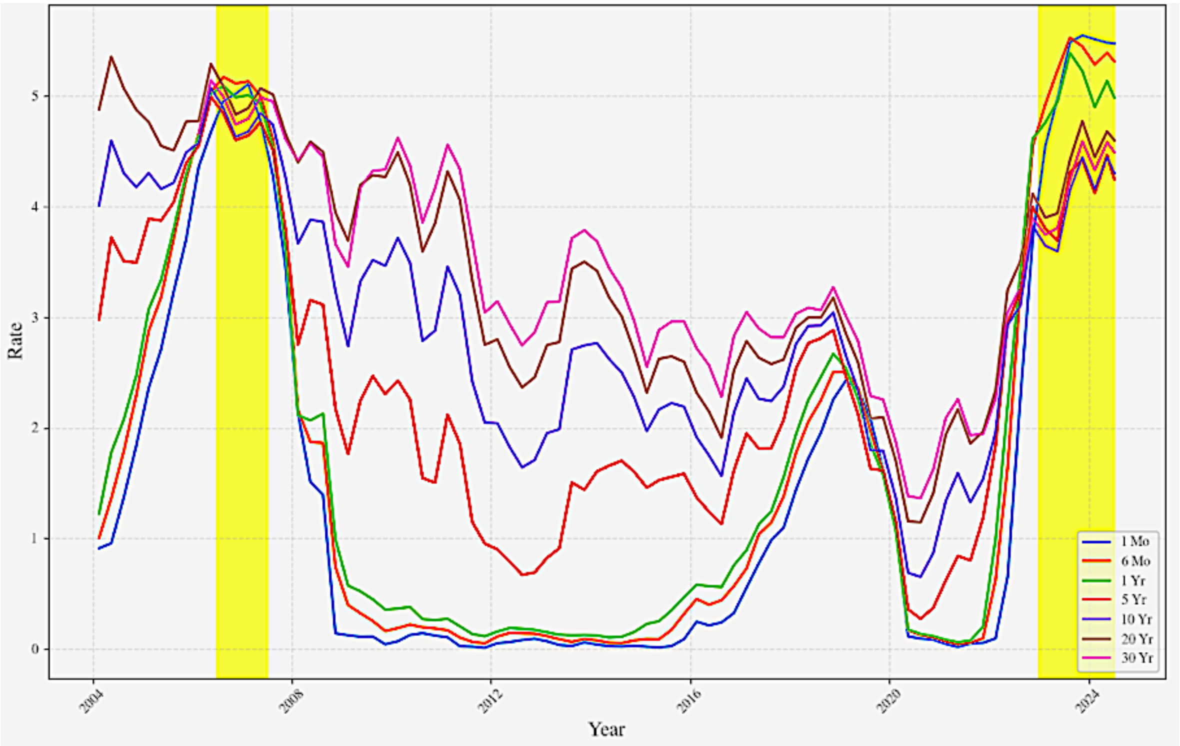
2.2.1 Yield Curve

The yield curve is a graphical representation of the interest rates on debt for a range of maturities, from short-term to long-term, at a given point in time. It shows the relationship between the yields (interest rates) of bonds and their time to maturity. The shape of the yield curve is a crucial indicator of market expectations regarding future interest rates and economic conditions.

Key types of yield curves include:

- **Normal Yield Curve**: Typically upward-sloping, indicating that longer-term interest rates are higher than short-term rates. This shape reflects expectations of economic growth and inflation.
- **Inverted Yield Curve**: Occurs when short-term interest rates are higher than long-term rates. An inverted curve can signal an anticipated economic slowdown or recession.
- **Flat Yield Curve**: Shows little difference between short-term and long-term interest rates, often occurring during transitions between economic cycles or periods of uncertainty.

In our analysis, the yield curve helps to understand how changes in interest rates across different maturities impact interest income. It provides insight into investor expectations and the cost of borrowing over different time horizons, which are crucial for assessing economic conditions and their effect on financial performance.



The above yield curve illustrates the relationship between interest rates of bonds with varying maturities over time. It provides valuable insights into the market's expectations for future interest rates and economic conditions. Typically, long-term interest rates (e.g., 10, 20, or 30-year rates) are higher than short-term rates (e.g., 1 or 2-month rates), reflecting the increased risk and uncertainty associated with longer-term investments. This upward slope of the yield curve generally indicates expectations of economic growth and moderate inflation. An inverted yield curve, where short-term interest rates exceed long-term rates, can signal an impending economic recession. The yellow areas on the curve denote specific periods when short-term rates were higher than long-term rates. Notably, the yield curve inverted just before the 2008 Financial Crisis and during the post-pandemic era, suggesting a potential economic slowdown during these times.

2.2.2 Yield Spread

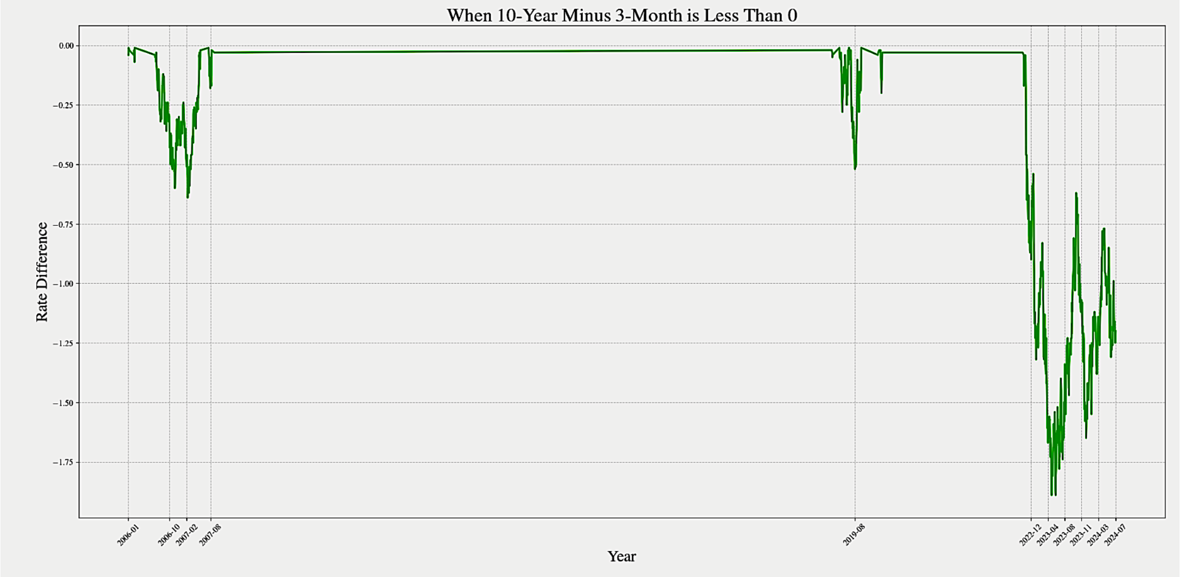
Yield spread refers to the difference in yields (interest rates) between two financial instruments, typically bonds or loans, with varying characteristics such as credit quality, maturity, or issuer type. It serves as a key indicator of relative value and risk between these instruments.

There are several types of yield spreads:

- **Credit Spread**: The difference between the yield of a corporate bond and a risk-free government bond of similar maturity. This spread reflects the credit risk premium required by investors for taking on additional credit risk associated with the corporate bond.
- **Maturity Spread**: The difference in yields between bonds with different maturities. This spread provides insight into investor expectations regarding future interest rates and economic conditions.
- **Term Spread**: The difference between yields on long-term and short-term government securities. This spread can signal expectations of future economic growth, inflation, and monetary policy changes.

In our dataset, which includes a total of 13 different interest rates, we analyze yield spreads by considering all possible combinations of these rates. Specifically, $\binom{13}{2}$, or 78 unique combinations, allow us to explore various spreads and their implications.

Yield spreads are crucial for assessing market conditions and investment opportunities. They offer insights into the perceived risk and return profiles of various securities, aiding investors in making informed decisions. In our analysis, yield spreads are used to understand market dynamics and to gauge the impact of interest rate changes on financial performance.



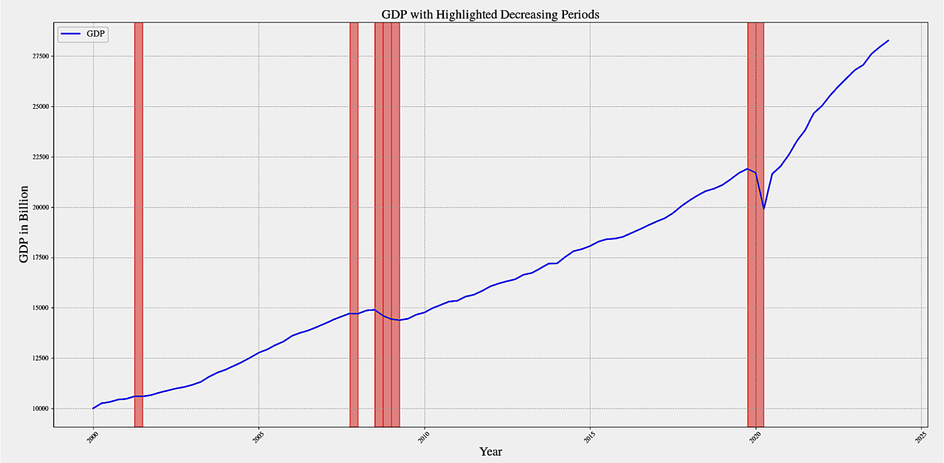
The plot generated displays the periods when the 10-year Treasury yield is lower than the 3-month Treasury yield. The trends clearly highlight key periods, such as the 2008 financial crisis, the COVID-19 pandemic, and the post-pandemic era. These events are marked by notable deviations in the yield curve.

2.2.3 Gross Domestic Product (GDP)

Gross Domestic Product (GDP) represents the total monetary value of all final goods and services produced within a country's borders during a specific period, typically a quarter. It is a critical indicator used to gauge the performance of an economy.

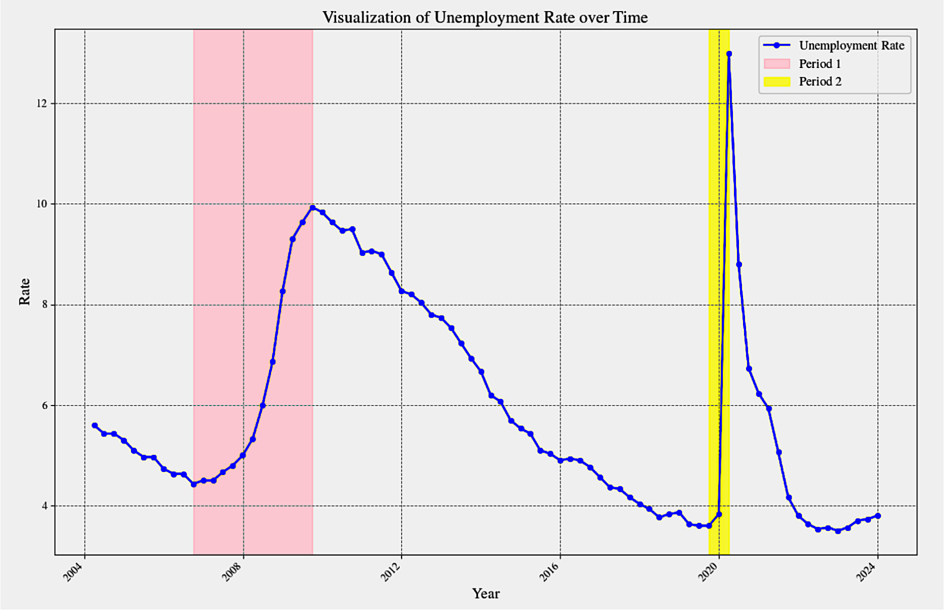
GDP Growth measures the rate at which a nation's economy is expanding or contracting over a specific period, typically a quarter or a year. It is calculated as the percentage change in GDP from one period to the next and serves as a crucial indicator of economic performance and health. Positive GDP growth indicates a growing economy, while negative growth may suggest a recession.

This plot below illustrates the Gross Domestic Product (GDP) over time, highlighting specific periods in red to indicate instances of GDP decline. Notable events associated with these declines include the 9/11 attacks, the 2008 Financial Crisis, and the COVID-19 pandemic. The purpose of this plot is to provide a clear visual representation of how significant external events can lead to substantial decreases in GDP.
Note: In our analysis, quarterly GDP data is used to understand economic trends and assess their impact on financial and investment decisions.



2.2.4 Unemployment Rate

The unemployment rate represents the percentage of individuals within the labor force who are actively seeking employment but are currently unemployed. It is a key indicator of labor market health and economic conditions.



The unemployment rate curve from 2004 to 2024 reveals key trends:

- **Financial Crisis (2008)**: A notable peak in 2009 reflects the rise in unemployment during the financial crisis.
- **COVID-19 Pandemic (2020)**: A significant spike in 2020 highlights the impact of the pandemic on job losses.
- **Recovery Trends**: Subsequent declines indicate gradual economic recovery and improvements in the labor market following each crisis.

The curve shown above effectively illustrates how major economic events influence unemployment and the labor market's recovery trajectory.

3 Modeling

In the modeling, when training and testing sets are needed, we randomly choose 80% of the dataset to be the training set, and the rest 20% to be the testing set.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

3.1 Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The primary goal of linear regression is to predict the value of the dependent variable based on the values of the independent variables.

The **dependent variable** Y is the outcome variable that the model aims to predict. The **independent variables** X are the input variables used to predict the dependent variable. A linear regression is called *simple* if there is only one independent variable.

The model we are using is **multiple linear regression**, which extends simple linear regression by using two or more independent variables to predict the dependent variable.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

- $X_1, X_2, \ldots, X_n$ are the independent variables.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients.

3.1.1 Assumptions of Linear Regression

The assumptions of linear regression are the following:

1. **Linearity**: The relationship between the dependent and independent variables is linear.
2. **Independence**: Input variables are independent of each other.
3. **Normality**: The residuals of the model are normally distributed.
4. **Homoscedasticity**: The variance of the residuals (errors) is constant across all levels of the independent variables.

3.1.2 Evaluating the Model

To access the performance of a linear regression model, various metrics can be used, such as:

- ***R*-squared** (R^2): Indicates the proportion of the variance in the dependence variable that is predictable from the independent variables.
- **Mean Squared Error (MSE)**: The average of the squared differences between the observed and predicted values.

3.1.3 Variable Choices

We choose several sets of variables to model the Interest Income and use R^2 and MSE to evaluate the model.
Set 1: 3-month, 1-year, 3-year, 10-year interest rates, unemployment rate, GDP.

- Result: R^2 Score = 0.436
- Interpretation: The data from FR Y-9C report varies from bank to bank. Set 1 only covers macroeconomic variables, so it is impossible to obtain a good prediction without considering any bank characteristics. In order to have a better result, we have to take account of variables corresponding to each particular bank.

Set 2: Lagged income, total assets.

- Result: R^2 Score = 0.790
- Interpretation: Interest income, lagged income and total assets are variables that change with banks, so it is understandable that we can get a much better prediction when using bank characteristics. How much can we improve? We need to combine macroeconomic indicators and bank characteristics together to apply the model again.

Set 3: 3-month, 10-year interest rates, unemployment rate, GDP, lagged income, total income.

- Result: R^2 Score = 0.815
- Interpretation: Combining everything so far together, we obtain the set of highest R^2 score. We always get higher R^2 score when adding new variables, but it does not make sense if we input everything we have. By comparing the R^2 score of Set 1 and Set 2, we know that to predict the interest income, bank characteristics are more effective, and adding macroeconomic indicators only improves R^2 score by 0.025.

Set 4: Lagged income.

- Result: R^2 Score = 0.790
- Interpretation: To see which of lagged income and total assets is a more important feature, we use one variable to model the interest income. It turns out that the R^2 score is improved by less than 0.0005 when adding total assets to lagged income. And the lagged income itself can model the interest income pretty well. Why is the total assets so unimportant? The main reason might be the normalization of interest income, in which step we divide the interest income of every bank by its total assets and get a ratio as the normalized interest income.

3.1.4 Limitations

In the model of linear regression, we assume that the relationship between the dependent variable and independent variables is linear and all input variables are independent of each other. These two facts are not always true. The macroeconomic indicators are usually related to each other. For example, when the interest rates are low, companies are more likely to take out loans to invest in capital projects, leading to increased production and higher GDP. When GDP is growing, businesses tend to hire more workers to meet increased demand for their products and services, which lowers the unemployment rates. Sometimes, interest rates are made lower to stimulate economic growth and reduce unemployment.

3.2 Regression Tree

A regression tree is a type of decision tree used to predict continuous outcomes. Unlike classification trees, which predict categorical outcomes, regression trees are used to predict numerical values. Regression trees are a part of a larger family of tree-based methods in machine learning and are particularly useful when there is a non-linear relationship between the dependent and independent variables. In the tree structure of a regression tree, each internal node represents a decision based on the value of an independent variables.

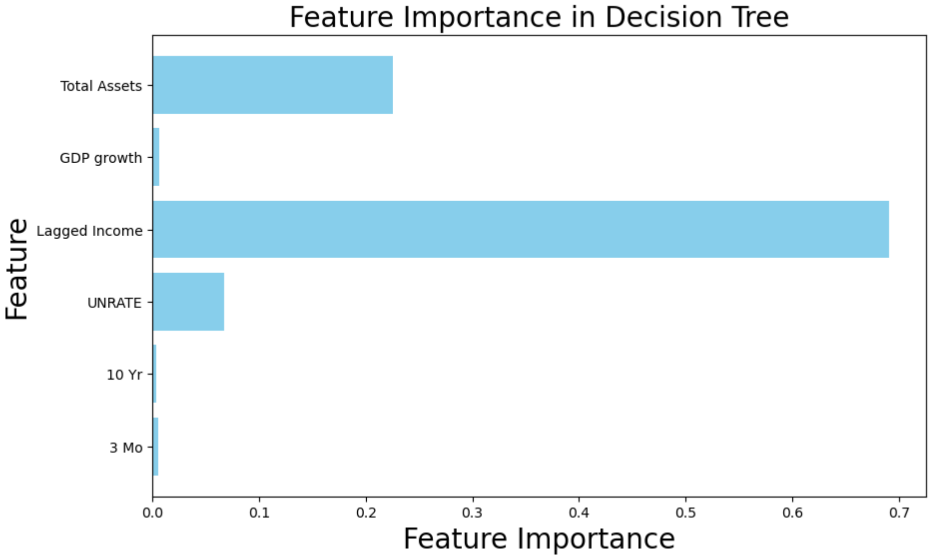
Advantages

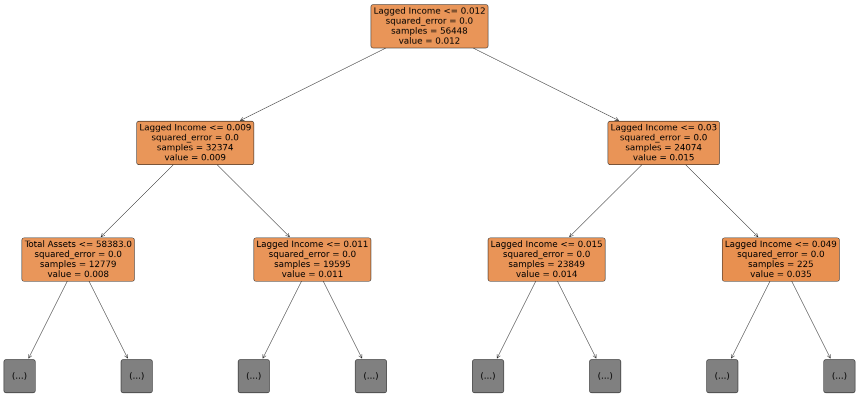
- **Easy to Understand and Interpret**: Regression trees are simple to visualize and interpret.
- **Non-parametric**: No assumptions about the underlying data distribution.
- **Handles Non-linearity**: Can model complex, non-linear relationships.

Disadvantages

- **Overfitting**: Trees can easily become too complex and overfit the training data, resulting in poor generalization to new data.
- **Instability**: Small changes in the data can lead to a completely different tree structure.
- **Bias-Variance Tradeoff**: High variance and low bias, meaning they can have high prediction errors.

In python, we can apply the regression tree to model by using the package *sklearn.tree* and get the evaluation of our model by using MSE and R^2. There are no assumptions about the underlying data distribution in the regression tree model, and we can use the function *feature_importances_* to see the importance of all input variables.





From the two pictures above, it is clear that the lagged income is the most important feature among all macroeconomic indicators and bank characteristics we take into consideration. In the regression tree, we see the first variable that is not the lagged income when the depth of regression tree is two. This concludes the same result which we obtain from the linear regression. The variable “Large Bank” is a categorical variable and is of the least importance. This is also because we normalize the interest income to be a ratio, which reduces the impact of total assets.

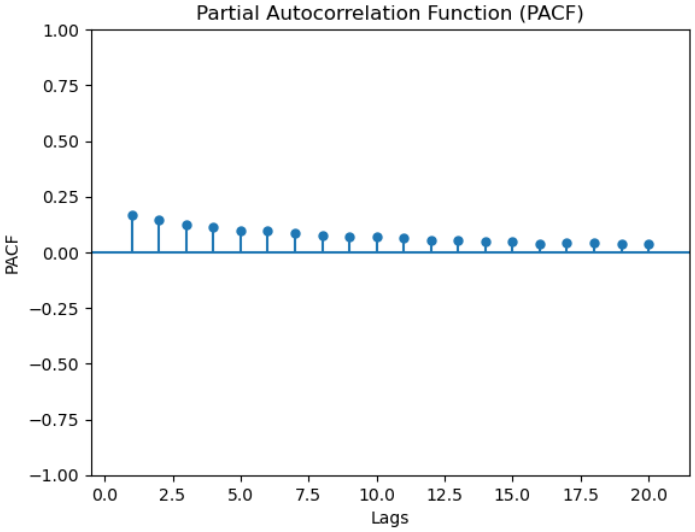
3.3 Autoregression

An autoregressive model is a type of statistical model used to analyze and forecast time series data. In an antoregressive model, the value of the variable at any point in time is assumed to be a linear function of its previous values plus a stochastic term (error term). Autoregressive models are widely used in time series analysis to understand the underlying structure of the data and to make forecasts.

The notation $AR(p)$ refers to an autoregressive model of order p. The order p indicates how many lagged values are used in the model. The $AR(p)$ model is defined as:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t$$

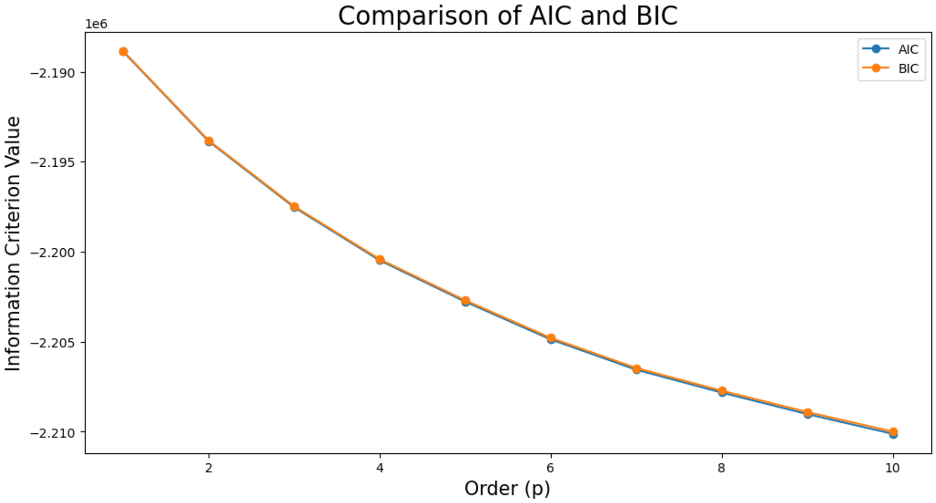
- X_t is the value of the time series at time t.
- $\phi_1, \phi_2, \ldots, \phi_t$ are the parameters of the model.
- ε_t is the white noise error term at time t, which is typically assumed to be normally distributed with mean zero and constant variance.



When building an AR model, the first step is to identify the order p. The order p of the autoregressive model can be identified using methods like the Partial Autocorrelation Function (PACF) plot or information criteria such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). First, we use PACF and obtain the following plot.

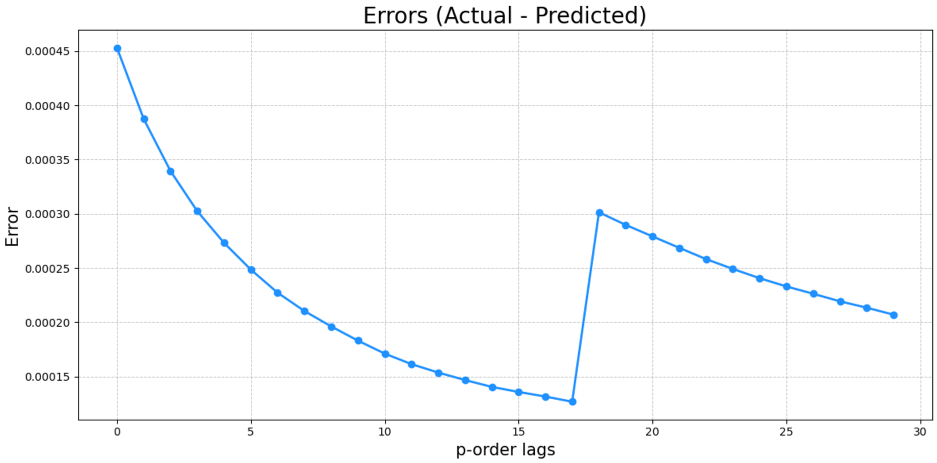
The plot above shows the decay pattern of PACF. As the order increases, the correlation values generally should decrease and approach to zero. A slow decay here suggests a persistent correlation, which a rapid decay will indicate a lack of long-term dependence. In the PACF plot, the correlation drops off after a certain lag and it suggests that the direct relationship exists only up to that lag.

Second, we compare the AIC and BIC for each different lag.



As seen in the plot above, AIC and BIC almost have the same values. The negative trend of them indicates that a model is better if it has higher order p.

We have that 80% of the data are used as the training set and the rest 20% are the test set. The following is the plot of errors, i.e., the absolute value of the difference between actual values and predicted values.



From this error plot, when the order $p = 17$, the error is minimum.

3.4 Random Forest

A random forest is an ensemble learning method used for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class (classification) or mean prediction (regression) of the individual trees. Random forests are known for their robustness, accuracy, and ability to handle large datasets with higher dimensionality.

Random forests combine the predictions of multiple models to produce a better overall prediction, and reduce the risk of overfitting compared to individual decision trees. In a random forest, multiple datasets, called bootstrap samples, from the original dataset is created. Then, on each bootstrap sample, a decision tree is trained. At each node in the tree, randomly select subset of features and choose the best split from this subset. For regression, we use the average of the predictions from all trees.

Advantages

- **High Accuracy:** Due to the ensemble nature, random forests usually provide more accurate predictions than individual decision tree.
- **Robustness:** Handles large datasets with higher dimensionality and is less prone to overfitting.
- **Feature Importance:** Provides an estimate of the importance of each feature in the prediction.

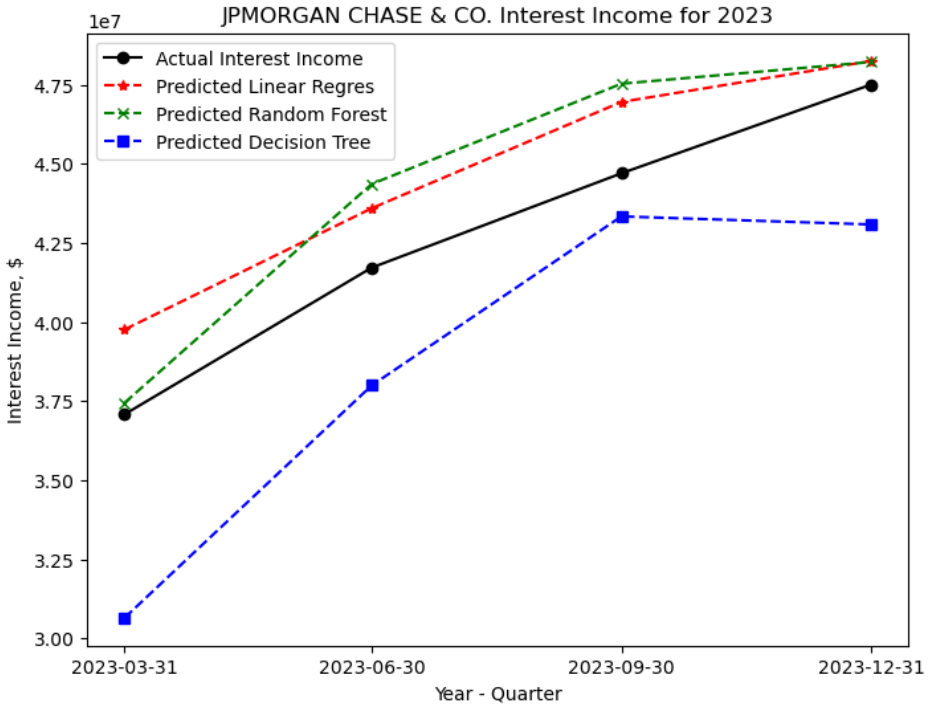
Disadvantages

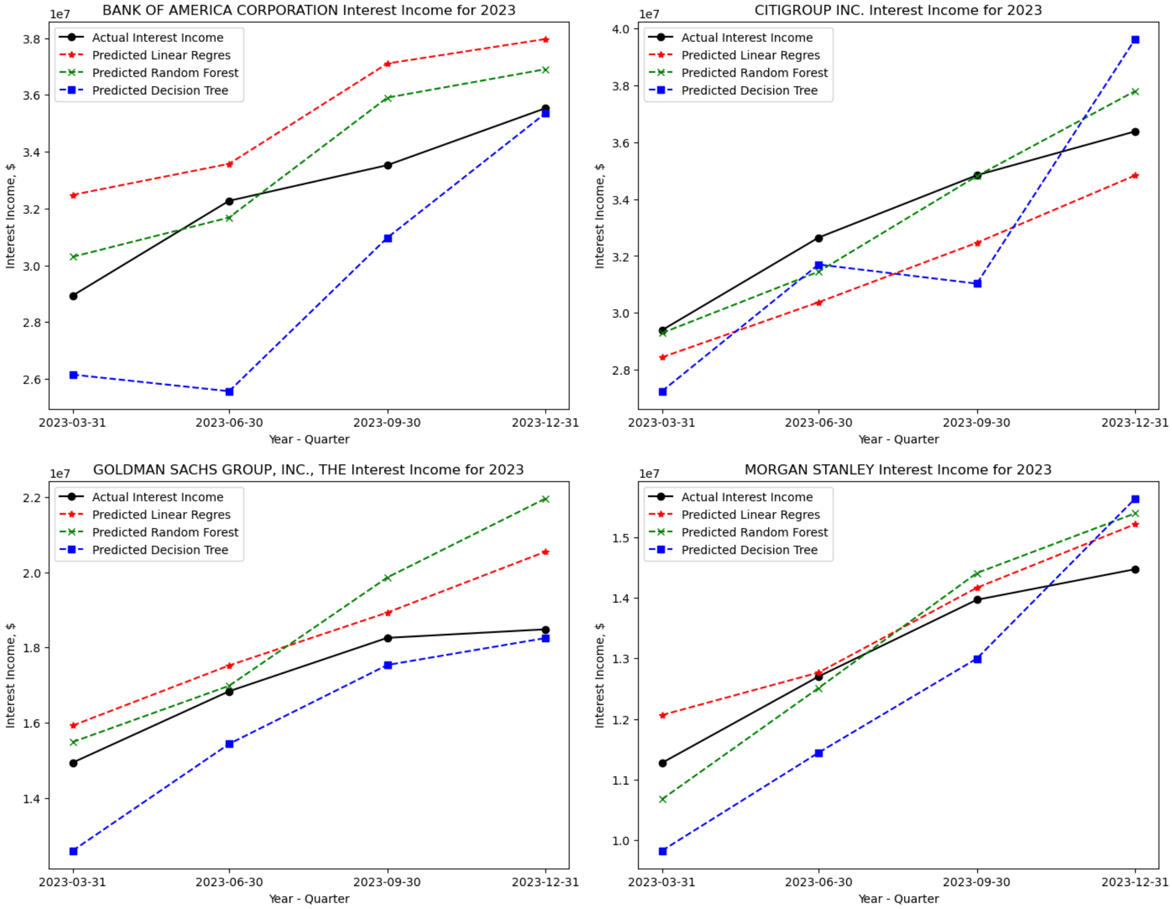
- **Complexity:** More complex and computationally intensive compared to single decision trees.
- **Interpretability:** Harder to interpret than individual decision trees.

Tuning the parameter *max_feature* (1, 2, ..., 7) by cross validation, we obtain that 3 features give the best R^2 score equal to 0.870, which is the best R^2 score over all models applied in this project.

4 Comparing Models

To compare the performance of different models, we use the data before 2023 as the training set and the data since 2023 to test the model. Repeating the models we have applied again, we can obtain models of similar performance with slight differences. To visualize the predicted values and actual values, we looked at the top five top banks of largest holdings in 2024: JP Morgan Chase, Bank of America, Citi Group, Goldman Sachs Group, and Morgan Stanley. Each plot below corresponds to one of these five top banks and shows predictions of interest income in dollars for 2023. The actual interest income (in dollars) plotted on the solid black line, the predicted linear regression plotted on the dashed red line, predicted random forest plotted on the dashed green line, and the predicted regression tree plotted on the dashed blue line.





We can draw some exciting features from the above plots, such as how all the model predictions for the top five banks follow the trend of the Actual Interest Income. We observe slight differences between the forecasts from the linear regression and random forest models compared to the actual interest income from the data; this may be due to the non-linear nature of the macroeconomic data. A remarkable aspect worth mentioning is the robustness demonstrated by the random forest model during this time interval. However, due to the complex nature of the decision tree model, its sensitivity to small data variations, and the high chance of overfitting, it resulted in poor performance compared to the other implemented models.

5 Conclusion

In conclusion, we have successfully modeled interest income using linear regression, regression trees, autoregression and random forests. From our comparison of the different models predicting the top 5 banks' interest incomes in the year 2023, we see that the random forest model is the best. We can come to the same conclusion when comparing the R^2 scores which was highest for the random forest model at 0.870.

For possible extension of this project, we suggest looking at modeling other components of PPNR and extending the dataset, such as using a longer period of time, and including other financial data reports such as the FR Y-14Q, or macroeconomic data that may affect the bank performance such as inflation rate, home price index, and stock market data.

Incorporating real-time economic and financial data could significantly improve the relevance and timeliness of forecasts. Additionally, developing hybrid models that combine change point detection with existing financial models has been shown to be beneficial; exploring this direction further could lead to better performance and results.

References

[1] U.S. Federal Reserve Board. Reporting forms: FR Y-9C. https://www.federalreserve.gov/apps/reportingforms/Report/Index/FR_Y-9C.

[2] U.S. Federal Reserve Board. Stress tests. https://www.federalreserve.gov/supervisionreg/stress-tests-capital-planning.htm.

[3] U.S. Federal Reserve. 2024 supervisory stress test methodology, 2024. https://www.federalreserve.gov/publications/files/2024-march-supervisory-stress-test-methodology.pdf.