**Time Series Forecasting for Quarterly Revenue Generation of State and Local Taxes**

**Group Members: -**

**Pranjali Desai**

**Prabhjot Gilard**

**Vinmathi Iyappan**

**Garima Vijay**

**Executive Summary: -**

This project embarks on a comprehensive exploration of regression modeling techniques applied to a time series dataset spanning from 2009 to 2022, focusing on quarterly revenue generation of State and Local Government Tax in the U.S. The dataset encapsulates quarterly estimates of tax revenue at the national and state levels, providing a nuanced perspective on fiscal trends measured in millions of dollars. The central objective is to evaluate the predictability of the data using a diverse array of regression models.

**Methodology:**

1. **Exploratory Data Analysis (EDA):**
   * Conduct an initial exploration of the dataset to identify trends, seasonality, or outliers.
   * Visualize data using plots and charts to gain insights.
2. **Models:**

* ARIMA
* Auto-ARIMA
* Quadratic
* Quadratic and Seasonality
* Exponential
* Model evaluation was based on the RMSE and MAPE accuracy metrics.

1. **Model-Specific Assessments:**
   * Evaluate the effectiveness of each model in capturing and forecasting tax revenue.
   * Fine-tune parameters for models such as quadratic trend, exponential trend, and ARIMA.

**Expected Outcomes:**

* A thorough analysis of predictability using a range of regression models.
* Insights into the effectiveness of different models in capturing and forecasting fiscal trends.
* Identification of potential non-linear patterns and seasonality within the dataset.

**Introduction: -**

The Quarterly Summary of State and Local Government Tax Revenue provides quarterly estimates of state and local government tax revenue at a national level, as well as detailed tax revenue data for individual states.  The information contained in this survey is the most current information available on a nationwide basis for government tax collections.

A graph of a number of people

Description automatically generated with medium confidenceWhile the state data records are ultimately from state government sources, the classification of taxes among the different categories is entirely the responsibility of the Census Bureau. Therefore, tax classification might not reflect the actual classification or presentation as requested by the various state government respondents.

**Eight steps of forecasting: -**

**Step 1: Define the Goal**

The primary goal of this analysis is to assess and enhance the predictability of monthly tax revenue values through the application of various modeling approaches. To initiate this process, we will employ traditional regression models to evaluate the linear relationship between time and monthly tax revenue. This foundational step aims to establish a baseline understanding of the data's behavior and identify potential patterns that can guide the subsequent modeling efforts. By delving into the traditional regression analysis, we set the stage for exploring more sophisticated modeling techniques, including quadratic trend models, exponential growth, or decay models, and ARIMA models with both manual and automated parameter identification. Through these steps, our overarching objective is to develop a robust forecasting model that effectively captures the underlying dynamics and seasonality inherent in the tax revenue data, thereby enabling more accurate predictions for future periods.

**Step 2: Get Data**

This report focuses on the time series dataset provided by the US Census Bureau recording the quarterly summary of state and local government tax revenue. The time period for the dataset ranges from 2009-2022 but for the purposes of this project the more recent timespans of 2009 to 2023 will be considered.

**Step 3: Explore and Visualize Series**

The data plots above show the respective upward trend and it’s a random walk. The autocorrelation of differenced tax data shows strongly negative correlation on lag 1. As the data is high correlated as the autocorrelation coefficients in all the lags are substantially higher at lag 1 than the horizontal threshold.

**A graph of a graph showing a number of different types of data

Description automatically generated with medium confidenceA graph with a line

Description automatically generated with medium confidenceA graph of tax time and numbers

Description automatically generated with medium confidenceStep 4: Data Preprocessing**

The original data from the Us Central Bureau for business and industry have the data from 1992 to 2023. But it is a quarterly data and the excel file has NA values from the year 1992 to 2009. So, we preprocessed the data and removed the incomplete data and next the dataset was created was from 2009 to 2023 quarterly data.

**Step 5: Partition the data.**

We created a data partition of 41 records for the training period and 15 records for the validation period. The purpose of this partitioning is to train the forecasting model on the train.ts data and then evaluate its performance on the unseen valid.ts data. This approach helps to assess the model's ability to generalize to new, unseen data, which is crucial for making reliable forecasts.

It's important to note that the partitioning of the data into training and validation sets is typically done in a way that preserves the temporal order of the observations, as time series data often exhibit autocorrelation and other temporal dependencies.

**Step 6 & 7: Apply Forecasting & Comparing Performance**

**1.Quadratic Trend Model Time Series: (Training/Validation):**

The choice of model depends on the characteristics observed in the time series plot. This data has quadratic trend hence applied this regression model.

**Summary for the training period**

A white text with black numbers and numbers

Description automatically generated

The regression model with quadratic trend contains a two independent variable: period index (t) and squared period index(t^2)

*y(t) = 254128.12 + 2204.74 t + 21.64 t^2*

According to the model summary, the regression model with quadratic trend is statistically significant. The coefficients for the intercept, trend (t) and quadratic trend (t2) variables are statistically significant (p-values for both coefficients are much lower than 0.05 or 0.01). The model has a good R-squared of 0.9856 (adj. R\_squared is 0.9848), which is a good fit for the training data.

The model can be used for forecasting.

A graph with a line and arrow

Description automatically generatedModel used to forecast in the validation period and the below in the accuracy score.

|  |  |
| --- | --- |
| **Time Series** | **Accuracy on the validation data** |
| **RMSE** | 45782.04 |
| **MAPE** | 8.521 |

**Entire dataset**

As the model has good accuracy score, the model is applied on the entire dataset and forecasted for the future 8 periods (2023-2024)

A screenshot of a computer

Description automatically generated**Summary for the historical data**

The regression model with quadratic trend contains a two independent variable: period index (t) and squared period index(t^2)

*y(t) = 267100 + 248.2 t + 69.16 t^2*

For the 11- year period, the model above represents a regression model with quadratic trend.

The coefficients for the intercept, quadratic trend (t2) variables are statistically significant (p-values for both coefficients are much lower than 0.05 or 0.01). The R-squared value is 0.9535, indicating that 95.35% of the variability in Revenue generated by Tax collection is explained by the model. The adjusted R-squared for the model is 95.17%.

A graph with a line and a blue line

Description automatically generatedThe model is used for forecasting the future period.

This graph shows overall a good fit for the historical data

|  |  |  |
| --- | --- | --- |
| TIME SERIES | VALIDATION PERIOD | ENTIRE DATSET |
| RMSE | 45782.04 | 15376.13 |
| MAPE | 8.521% | 2.678% |

**Accuracy score**

**2.Quadratic Trend and Seasonality Model Time Series: (Training/Validation):**

**Summary for the training period**

The regression model with quadratic trend and seasonality contains 5 independent variables: period index (t) (t^2) and 3 seasonal dummy variables for Q2 (season2 – D2), Q3 (season3 – D3) and Q4 (season4 – D4).

The model’s equation is:

*Y(t) = 255120.913 + 2224.7 t + + 21.121 t^2 – 2229.491 D2 – 1043.761 D3 – 1290.872 D4*

According to the model summary, the regression model with quadratic trend and seasonality is statistically significant. The coefficients for the intercept, trend (t) and quadratic trend (t2) variables are statistically significant (p-values for both coefficients are much lower than 0.05 or 0.01). The model has a good R-squared of 0.9851 (adj. R\_squared is 0.9841), which is a good fit for the training data.

The model can be used for forecasting.

Model used to forecast in the validation period and the below in the accuracy score.

A graph with a line going up

Description automatically generated

|  |  |
| --- | --- |
| **Time Series** | **Accuracy on the validation data** |
| **RMSE** | 46069.06 |
| **MAPE** | 8.579 |

**Entire dataset**

As the model has good accuracy score, the model is applied on the entire dataset and forecasted for the future 8 periods (2023-2024)

**Summary for the historical data**

A screenshot of a computer

Description automatically generated

The model’s equation is:

*Y(t) = 268759.02 + 248.49 t + + 69.14 t^2 – 5377.60 D2 + 517.43 D3 – 1574.18 D4*

For the 11- year period, the model above represents a regression model with quadratic trend and seasonality.

The coefficients for the intercept, quadratic trend (t2) variables are statistically significant (p-values for both coefficients are much lower than 0.05 or 0.01). The R-squared value is 0.9546, indicating that 95.46% of the variability in Revenue generated by Tax collection is explained by the model. The adjusted R-squared for the model is 95%.

A graph with a line going up

Description automatically generatedThe model is used for forecasting the future period.

This graph shows overall a good fit for the historical data

|  |  |  |
| --- | --- | --- |
| TIME SERIES | VALIDATION PERIOD | ENTIRE DATSET |
| RMSE | 46069.06 | 15201.87 |
| MAPE | 8.579 | 2.678 |

**Accuracy score**

A white sheet with black numbers and a white background

Description automatically generated with medium confidence**3.Exponential Trend Model Time Series: (Training/Validation):Summary for the training period**

The regression model with exponential trend contains a exponential trend as an independent variable

The model’s equation is:

*Ln(t) = 12.44 + 0.009952t*

According to the model summary, the regression model with exponential trend is statistically significant. The coefficients for the intercept and trend (t) are statistically significant (p-values for both coefficients are much lower than 0.05 or 0.01). The model has a good R-squared of 0.9855 (adj. R\_squared is 0.9851), which is a good fit for the training data.

The model can be used for forecasting.

A graph with a line and arrow

Description automatically generated

Model used to forecast in the validation period and the below in the accuracy score.

|  |  |
| --- | --- |
| **Time Series** | **Accuracy on the validation data** |
| **RMSE** | 48677.99 |
| **MAPE** | 9.177 |

**Entire dataset**

As the model has good accuracy score, the model is applied on the entire dataset and forecasted for the future 8 periods (2023-2024)

**Summary for the historical data**

A screenshot of a computer code

Description automatically generated

The model’s equation is:

*Ln(t) = 12.40 + 0.011 t*

For the 11- year period, the model above represents a regression model with exponential trend

The coefficients for the intercept, exponential trend variables are statistically significant (p-values for both coefficients are much lower than 0.05 or 0.01). The R-squared value is 0.945, indicating that 94.5% of the variability in Revenue generated by Tax collection is explained by the model. The adjusted R-squared for the model is 94%.

The model is used for forecasting the future period.

A graph with a line and a blue line

Description automatically generated

|  |  |  |
| --- | --- | --- |
| TIME SERIES | VALIDATION PERIOD | ENTIRE DATSET |
| RMSE | 48677.99 | 18680.78 |
| MAPE | 9.177 | 3.056 |

**Accuracy**

**4.ARIMA MODEL**

We generated an ARIMA model with (1,1,1)(1,1,1) parameters using the Arima() function.

**ARIMA for Training & Validation Data**

A screenshot of a computer

Description automatically generated

The model’s equation is:

*yt - yt-1 = 0.3510 (yt-1 -yt-2) - 0.8123 εt-1 - 0.4829 (yt-1 -yt-5) - 0.145 ρt-1*

This is a seasonal ARIMA model, (1,1,1)(1,1,1)[4], with the following parameters:  
• p = 1, order 1 autoregressive model  
• d = 1, first differencing  
• q = 1, order 1 moving average MA(1) for error lags  
• P = 1, order 1autoregressive model for the seasonal part  
• D = 1, first differencing for the seasonal part  
• Q = 1 order 1 moving average model for the seasonal part’s error lags  
• m = 4, for the quarterly seasonality.

|  |  |
| --- | --- |
| Time Series | Training &Validation Period ARIMA MODEL |
| RMSE | 5179.978 |
| MAPE | 1.220% |

AR(1) for seasonality- Order 1 indicates there is dependence on the previous seasonal observation.  
Differencing (d)- Order 1 suggests that the time series has been differenced once to remove a linear trend.

MA(1), moving average component- order 1 implies that the current observation is influenced by the past observation's error.

A screenshot of a computer

Description automatically generatedTo summarize, the ARIMA(1,1,1)(1,1,1)[4] model is a more complex model it includes autoregressive, differencing, and moving average terms for both the original and seasonal series.

A graph with lines and text

Description automatically generated

As all the auto-correlation values lies inside the boundary of confidence for residuals  
which is a positive outcome, suggesting that the ARIMA model adequately captures the temporal patterns in the data for training and validation period.

A graph of a graph showing the growth of training

Description automatically generated

ARIMA Model Forecasting on the Entire Dataset

The model’s equation is:

*yt - yt-1 = -0.2578 (yt-1 -yt-2) - 0.3923 εt-1 - 0.0702 (yt-1 -yt-5) – 0.8703 ρt-1*

A screenshot of a computer

Description automatically generatedThis is a seasonal ARIMA model, (1,1,1)(1,1,1)[4], with the following parameters:  
• p = 1, order 1 autoregressive model  
• d = 1, first differencing  
• q = 1, order 1 moving average MA(1) for error lags  
• P = 1, order 1autoregressive model for the seasonal part  
• D = 1, first differencing for the seasonal part  
• Q = 1 order 1 moving average model for the seasonal part’s error lags  
• m = 4, for the quarterly seasonality.

A graph with lines and text

Description automatically generated

A screenshot of a computer screen

Description automatically generated

* The point forecasts provide the model's best estimate for the expected values of the time series at each specified quarter.
* For instance, in 2023 Q1, the model predicts a point forecast of 510,193.3. This represents the anticipated value for the given quarter.
* Similarly, each subsequent quarter in 2023 and 2024 has its corresponding point forecast.
* As all the auto-correlation values lies in side the boundary of confidence for residuals  
  which is a positive outcome, suggesting that the ARIMA model adequately captures the temporal patterns in the data for entire dataset as well.

|  |  |  |
| --- | --- | --- |
| **TIME SERIES** | **TRAINING &VALIDATION PERIOD ARIMA MODEL** | **ENTIRE DATSET ARIMA MODEL** |
| **RMSE** | 5179.97 | 16504.35 |
| **MAPE** | 1.220 | 2.295 |

A graph with a line going up

Description automatically generated

The graph shows a good fit for the historical data.

**4. AUTO ARIMA MODEL**

We generated an optimal ARIMA model with automatic selection of (p,d,q)(P,D,Q) parameters using the Arima() function.

**Auto ARIMA for Training & Validation Data**

A black text with numbers

Description automatically generated with medium confidence

ARIMA(0,1,0) – this model does not use auto regressive model.

The model’s equation is:

*y(t) -y(t-1) = 4482.37*

This is a ARIMA model, (0,1,) with the following parameters:  
• p = 0, no autoregressive model  
• d = 1, first differencing  
• q = 1, no moving average MA(1) for error lags.

|  |  |
| --- | --- |
| Time Series | Training &Validation Period ARIMA MODEL |
| RMSE | 5218.368 |
| MAPE | 1.24 |

A graph with blue lines

Description automatically generated  
Differencing (d)- Order 1 suggests that the time series has been differenced once to remove a linear trend.

No autocorrelation exists in the training dataset.

A screenshot of a computer

Description automatically generated

A graph with a line going up

Description automatically generated

The graph shows the good fit for training and slightly underestimated for validation period

**Entire dataset**

Summary

A black text on a white background

Description automatically generated

ARIMA(0,1,1) – this model does not use auto regressive model.

The model’s equation is:

*yt - yt-1 = 4482.37 - 0.5653 εt-1*

This is a ARIMA model, (0,1,1) with the following parameters:  
•p=0, noautoregressivemodel  
•d=1,first differencing  
• q = 1, order 1 moving average MA(1) for error lags.

A graph with blue lines

Description automatically generated

A graph with a line going up

Description automatically generatedAs all the auto-correlation values lies inside the boundary of confidence for residuals  
which is a positive outcome, suggesting that the ARIMA model adequately captures the temporal patterns in the data for entire dataset as well.

|  |  |  |
| --- | --- | --- |
| **TIME SERIES** | **TRAINING &VALIDATION PERIOD ARIMA MODEL** | **ENTIRE DATSET ARIMA MODEL** |
| **RMSE** | 5218.368 | 16316.58 |
| **MAPE** | 1.24 | 2.414 |

**Step 8: -**

**A screenshot of a computer

Description automatically generatedCompare the accuracy of all the models.**

**Conclusion: -**

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **MAPE** |
| ARIMA Model with seasonality | 6504.35 | 2.30% |
| Auto ARIMA | 6316.58 | 2.41% |
| Quadratic Regression model | 15376.13 | 2.58% |
| Quadratic model with seasonality | 15201.87 | 2.69% |
| Exponential model | 18680.78 | 3.05% |
| Naive | 20251.23 | 2.61% |
| Seasonal Naive | 29619.99 | 5.48% |

The comprehensive analysis of diverse regression models for predicting State and Local Government Tax revenue indicates that the ARIMA model with seasonality emerges as the most suitable choice for this specific time series analysis project.

The project, aimed at providing valuable insights into fiscal trends, offers a systematic exploration of various models. The ARIMA model's strong performance, as evidenced by its competitive ranking in MAPE: 2.30% metrics having lowest error values as compared to all models and also in combination of RMSE and MAPE it stands out.  
The ARIMA model underscores its reliability in capturing and predicting underlying patterns in tax revenue data.

The practical applications of these findings are substantial, especially in the realm of informed decision-making for policymakers and stakeholders. The ARIMA model's accuracy and effectiveness make it a valuable tool for financial planning and resource allocation within the context of State and Local Government fiscal management in the U.S.

By choosing ARIMA as the preferred model, this project lays the groundwork for a robust framework that can enhance the precision and reliability of predictions, ultimately supporting prudent decision-making processes in the dynamic landscape of government finances.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated**A table with numbers and a number of numbers

Description automatically generated**Comparing the point forecast for ARIMA and Auto ARIMA with the original values which were available in for the first three quarters of 2023, we can observe that these values approximately being similar validates that the autoregressive models are the best fit predict State and Local Government Tax revenue.  
A table with numbers and a number of numbers

Description automatically generatedTop of Form