Adrian Vasquez

Forecasting Report

ISDS 599B

12-27-23

**Executive Summary**

This report presents the findings of my independent study on advanced forecasting using RStudio libraries Fable and FPP3. The study focuses on forecasting the sales of an Ecuadorian grocery store similar to Walmart, utilizing a public dataset from Kaggle.

**Data:**

The data for this analysis is available on Kaggle, a public platform for sharing and exploring data sets. The data can be accessed at the following link: <https://www.kaggle.com/c/store-sales-time-series-forecasting/data>.

Since the data is publicly available on Kaggle, there are no legal or privacy concerns associated with its use.

The data is provided in several CSV files, offering a comprehensive set of information relevant to the forecasting analysis. These files include:

* Sales data: This contains information on store number, department, promotions, and sales.
* Holiday data: This provides both local and regional information on holidays that may impact sales.
* Store location data: This details the location of each store.
* Oil price data: This provides information on oil prices, potentially influencing the economy and sales.

To effectively analyze the data and capture all the relevant information, these separate data sets will need to be merged into a single, comprehensive data set. This will facilitate a more holistic understanding of the factors affecting sales and enable the development of more accurate and reliable forecasts. However, for this project I am not going to use store location data, and oil price data

**Objective**

My objective is to learn how to forecast using r, with a special emphasis on how to forecast hierarchical time series. Hierarchical forecasting involves creating a series of forecasts at different levels of aggregation. For example, you might forecast individual product sales at the store level, then combine those forecasts to create a forecast for total sales across all stores. The data I have has 54 stores and 34 different departs such as magazines, beauty, healthcare, grocery, and automotive to name a few.

1. My goal is to create 1800 + forecast models at every level of the hierarchy.
2. Forecasting using all the method explain in the following book Forecasting: Practices and Principles by Rob Hyndman and George Athanasopoulos
3. Evaluate the model based on AIC, RMSE, and MAPE.

**Data Visualization**

**A graph with a line graph

Description automatically generated** A graph of sales and shopping

Description automatically generated with medium confidence

**A graph of a number of months

Description automatically generated with medium confidence  
Forecasting Techniques:**

For this report we will perform and evaluate several forecasting techniques such as Exponential smoothing, Regression, ARIMA, and Decomposition. Fore the purpose of learning I will forecast using a single store(store3) and a single department(grocery 1).

**Naive Forecasting**

Concept: The most basic forecasting method, assuming the future value will be identical to the most recent observed value.

* Formula: F\_{t+1} = Y\_t, where F\_{t+1} is the forecast for the next period and Y\_t is the actual value in the current period.
* Example: If sales last month were 100 units, the naïve forecast for next month would also be 100 units.
* Pros: Simple to understand and implement, requires no model fitting.
* Cons: Often inaccurate, especially for data with trends or seasonality.
* Should only be used as a benchmark or if the data does not have any seasonality or trend.

A graph on a grid

Description automatically generated

A close-up of numbers

Description automatically generated

A screenshot of a graph

Description automatically generated

**Decomposition**:

Classical Decomposition (Additive and Multiplicative):

* This method separates the time series into trend, seasonality, and residual components.
* Additive decomposition assumes the components are added together, while multiplicative assumes they are multiplied.
* Both methods are suitable for stationary time series and can provide insights into the underlying trends and seasonal patterns.

A graph of sales and seasonal data

Description automatically generated with medium confidence

STL Decomposition:

* This non-parametric method is robust to outliers and can handle non-linear trends and seasonality.
* It is flexible and can be applied to various time series, including daily, weekly, and monthly data.
* However, it requires more expertise than classical decomposition and may not be suitable for very long series.

For the purpose of this book the only taught how to forecast using STL decomposition.

A graph showing the us retail employment rate

Description automatically generated

A close-up of a computer code

Description automatically generated

Census Decomposition Methods:

* These methods are designed for monthly and quarterly data and are known for their robustness and automation.
* X-11: This method is highly robust to outliers and automatically chooses appropriate trend and seasonal models. It is suitable for long time series but lacks confidence intervals.
* X-12ARIMA: This method combines X-11 with ARIMA modeling for improved accuracy and includes confidence intervals.
* X-13ARIMA-SEATS: This is the latest version that incorporates additional features like trading day adjustments and holiday effects.

A graph of sales

Description automatically generated

A graph with blue lines and white text

Description automatically generated

**Exponetial Smoothing**: Brown, Holt, and Winters

Exponential smoothing is a popular forecasting technique that uses weighted averages of past observations to predict future values. Different methods of exponential smoothing can be used for different types of time series data.

1. Brown's simple exponential smoothing:

* This is the simplest exponential smoothing method and is suitable for forecasting stationary time series with no trend or seasonality.
* It assigns weights to past observations with the most recent observation receiving the highest weight.
* The weight assigned to each observation decays exponentially over time, giving less weight to older observations.

2. Holt's linear exponential smoothing:

* This method extends Brown's method to include a tre
* nd factor.
* It estimates both the level and the trend of the time series and uses them to make forecasts.
* The formula for Holt's linear exponential smoothing is:

3. Winters' triple exponential smoothing:

* This method extends Holt's method to include seasonality.
* It estimates the level, trend, and seasonal factors and uses them to make forecasts.
* The formula for Winters' triple exponential smoothing is:

Choosing the right exponential smoothing method:

* Use Brown's simple exponential smoothing if your data is stationary and has no trend or seasonality.
* Use Holt's linear exponential smoothing if your data has a trend but no seasonality.
* Use Winters' triple exponential smoothing if your data has both a trend and seasonality.
* The values of the smoothing constants (α, β, and γ) need to be chosen carefully. You can experiment with different values to find the ones that produce the most accurate forecasts.
* There are variations of these methods that can be used for more complex time series data, such as those with multiple seasonality’s or missing values.

To help determine what exponential smoothing model we first need to visualize our data. From there we need to determine if our data has trend, exponential trend, dampened trend, and if it has multiplicative or additive seasonality:

**A graph with a line graph

Description automatically generated**

Our data displays a positive linear trend with no apparent seasonality. This suggests a simple linear trend model without seasonal terms might be suitable for forecasting. However, relying solely on intuition can be risky.

Luckily Fable has a powerful forecasting package that offers a built-in function to automatically select the best model and its optimal smoothing parameters. It employs a sophisticated strategy combining:

* Grid Search: Exploring a predefined range of parameter values for various models.
* Maximum Likelihood Estimation: Finding the parameter values that best explain the observed data.
* Kalman Filter: Adaptively adjusting parameters based on the evolving data patterns .

Ultimately fable chooses the model with the lowest AIC (Akaike Information Criterion). AIC not only assesses how well the model fits the data but penalizes overly complex models, encouraging a balance between accuracy and simplicity.

To determine which model is best I use the fable automation process verse my intuition.

A graph of a graph showing a line graph

Description automatically generated with medium confidence

> accuracy(MAM)

# A tibble: 1 × 10

.model .type ME RMSE MAE MPE MAPE MASE RMSSE ACF1

*<chr>* *<chr>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 AN Training -5988. 31737. 19703. -4.47 9.57 0.762 0.985 0.124

report(MAM)

Series: Monthly\_sales

Model: ETS(M,A,N)

Smoothing parameters:

alpha = 0.05575555

beta = 0.01195125

Initial states:

l[0] b[0]

163437.8 2916.384

sigma^2: 0.0163

AIC AICc BIC

1385.778 1386.978 1395.905

**Fables automated EST model selection:**

A graph showing the fall of a company

Description automatically generated

> accuracy(FableEST)

# A tibble: 1 × 10

.model .type ME RMSE MAE MPE MAPE MASE RMSSE ACF1

*<chr>* *<chr>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 FABleEST Training -478. 29559. 16973. -2.16 8.14 0.656 0.917 0.0807

report(FableEST)

Series: Monthly\_sales

Model: ETS(M,Ad,N)

Smoothing parameters:

alpha = 0.0001002222

beta = 0.0001000058

phi = 0.9609186

Initial states:

l[0] b[0]

159739.2 4838.061

sigma^2: 0.0152

AIC AICc BIC

1380.583 1382.298 1392.735

In a comparison of the automated EST model and my domain knowledge of the exponential forecasting models, Fable's automated ETS selection process favored the Holt's Linear Multiplicative Error model, while my initial choice was the Holt's Linear Trend model. To evaluate their performance, I assessed RMSE, MAPE, and AIC.

Key Findings:

* Holt's Linear Trend Model (My Selection):
  + AIC: 1385.778
  + MAPE: 9.57%
  + RMSE: 31737
* Holt's Linear Multiplicative Error Model (Automated ETS Selection):
  + AIC: 1380
  + MAPE: 8.14%
  + RMSE: 29559

Conclusion:

* The automated ETS selection outperformed my initial choice across all three metrics:
  + Lower AIC (1380 vs. 1385.778) suggests a better balance between model fit and complexity.
  + Lower MAPE (8.14% vs. 9.57%) indicates more accurate forecasts, with an average error of only 8.14%.
  + Lower RMSE (29559 vs. 31737) signifies smaller overall forecast errors.

Therefore, based on these results, the Holt's Linear Multiplicative Error model is deemed superior for generating more accurate and reliable forecasts in this instance.

**ARIMA :**

ARIMA, which stands for Autoregressive Integrated Moving Average, is a powerful statistical method used for forecasting time series data. It combines three key components to analyze and predict future values:

1. Autoregressive (AR): This component models the dependence of the current value on its past values. It uses a weighted linear combination of past values to predict the current value. The weights are determined by the model parameters and reflect the strength and decay of the relationship between past and future observations.

2. Integrated (I): This component removes non-stationarity from the data. Non-stationarity refers to situations where the mean, variance, or autocorrelation of the data changes over time. To achieve stationarity, ARIMA models use differencing, which involves subtracting the previous value from the current value.

3. Moving Average (MA): This component models the dependence of the current value on its past forecast errors. It uses a weighted linear combination of past forecast errors to adjust the predicted value. The weights represent the impact of past errors on the current prediction, helping to correct for any systematic biases in the forecasts.

ARIMA model order:

ARIMA models are identified by their order, denoted by three numbers: p, d, and q:

* p: Order of the AR term (number of past values used for prediction)
* d: Order of differencing (number of times the data needs differencing to achieve stationarity)
* q: Order of the MA term (number of past forecast errors used for correction)

For example, an ARIMA(1,1,1) model includes one AR term, one differencing step, and one MA term.

Working process:

1. Data preprocessing: Analyze the time series data to identify any patterns, trends, and seasonality.

A graph showing a line graph

Description automatically generated

The data has a trend component but no seasonality so a single differencing should make the data stationary.

1. Stationarity check: Evaluate if the data is stationary. If not, apply differencing to achieve stationarity.

A graph with lines on a gray background

Description automatically generated

A graph with lines and numbers

Description automatically generated

|  |
| --- |
| monthly\_stor3\_grocery1 |>  + features(Monthly\_sales, unitroot\_ndiffs)  # A tibble: 1 × 1  ndiffs  *<int>*  1 1  > ###seasonal differencing  > monthly\_stor3\_grocery1 |>  + features(Monthly\_sales, unitroot\_nsdiffs)  # A tibble: 1 × 1  nsdiffs  *<int>*  1 0 |
|  |
| |  | | --- | |  | |

According to the unit root test only first order differencing is necessary, and the ACF of original data and differenced data shows that first order differencing is sufficient.

The automatic model identification determined the appropriate order of (p, d, q) of the ARIMA model based on analysis of the autocorrelation function (ACF) and partial autocorrelation function (PACF).

Using Fable The (p,d,q) is as Follows: ARIMA(0,1,1)(1,0,0)[12]

|  |
| --- |
| Series: Monthly\_sales  Model: ARIMA(0,1,1)(1,0,0)[12]  Coefficients:  ma1 sar1  -0.7159 0.6940  s.e. 0.1605 0.1082  sigma^2 estimated as 644705659: log likelihood=-639.13  AIC=1284.26 AICc=1284.73 BIC=1290.28 |
|  |
| |  | | --- | |  | |

1. Model diagnostics: Evaluate the performance of the model by analyzing residuals and comparing forecasts with actual values.

|  |
| --- |
|  |
|  |
| |  | | --- | | > | |  | |

A close up of numbers

Description automatically generated

1. Forecasting: Use the estimated model parameters to generate forecasts for future time periods.
2. Arima\_fit |> forecast(h=12)
3. # A fable: 12 x 4 [1M]
4. # Key: .model [1]
5. .model Monthly\_date Monthly\_sales .mean
6. *<chr>* *<mth>* *<dist>* *<dbl>*
7. 1 ARIMA(Monthly\_sales) 2017 Sep N(222526, 6.4e+08) 222526.
8. 2 ARIMA(Monthly\_sales) 2017 Oct N(237057, 7e+08) 237057.
9. 3 ARIMA(Monthly\_sales) 2017 Nov N(240122, 7.5e+08) 240122.
10. 4 ARIMA(Monthly\_sales) 2017 Dec N(289293, 8e+08) 289293.
11. 5 ARIMA(Monthly\_sales) 2018 Jan N(245334, 8.5e+08) 245334.
12. 6 ARIMA(Monthly\_sales) 2018 Feb N(222826, 9e+08) 222826.
13. 7 ARIMA(Monthly\_sales) 2018 Mar N(248652, 9.6e+08) 248652.
14. 8 ARIMA(Monthly\_sales) 2018 Apr N(242420, 1e+09) 242420.
15. 9 ARIMA(Monthly\_sales) 2018 May N(242318, 1.1e+09) 242318.
16. 10 ARIMA(Monthly\_sales) 2018 Jun N(235148, 1.1e+09) 235148.
17. 11 ARIMA(Monthly\_sales) 2018 Jul N(234386, 1.2e+09) 234386.
18. 12 ARIMA(Monthly\_sales) 2018 Aug N(132351, 1.2e+09) 132351.

A graph of a graph

Description automatically generated

Strengths of ARIMA:

* Flexibility: ARIMA models can handle various data patterns and seasonality by adjusting the model order.
* Interpretability: The model parameters offer insights into the relationships between past and future values.
* Widely used: ARIMA is a well-established and widely used forecasting technique with readily available software implementations.

Limitations of ARIMA:

* Stationarity requirement: Non-stationary data requires differencing, which can discard valuable information.
* Limited ability to detect complex patterns: ARIMA might struggle with highly non-linear or complex patterns in the data.
* Model selection and parameter estimation can be complex: Choosing the correct model order and estimating parameters can be challenging for beginners.

Overall, ARIMA remains a powerful and versatile tool for forecasting time series data. Its ability to capture linear relationships, handle seasonality, and offer interpretable results makes it a valuable choice for various forecasting tasks.

**Regression**:

Regression models predict the value of a dependent variable based on the values of one or more independent variables. In forecasting, the dependent variable is the future value of the time series, and the independent variables can be past values of the series, external factors, or other relevant information.

There are various types of regression models used in forecasting, such as linear regression, polynomial regression, dynamic regression, and harmonic regression for daily or hourly data.

In order to run regression the data need to meet the following four assumptions:

1. Linearity:

* The relationship between the variables you're studying can be plotted as a straight line. If it's curved, you might need to transform variables or use a different model.

2. Independent Errors:

* Errors (the differences between actual and predicted values) don't influence each other. This is especially important in time series data, where errors often tend to be correlated over time.

3. Homoscedasticity:

* Errors have a constant variance across all levels of the independent variables. If they don't, your confidence intervals and predictions might be unreliable.

4. Normality of Errors:

* Errors are normally distributed. While not strictly necessary for estimation, it's important for hypothesis testing and constructing confidence intervals.

5. No Multicollinearity:

* Independent variables aren't too highly correlated with each other. This can make it hard to isolate the effect of each variable on the outcome.

Checking Assumptions Matters:

* It's important to check these assumptions to ensure the validity of your results and the reliability of your conclusions.
* If assumptions are violated, you might need to adjust your model or use alternative techniques.

To ensure are model meet this assumption we need to check a plot of the residuals and a plot of the of the sales data.

**A graph with a line graph

Description automatically generated**

A screenshot of a graph

Description automatically generated

The line plot of the monthly data clearly exhibits a positive linear relationship, satisfying the linearity assumption of regression analysis.

To assess the assumptions of independence, homoscedasticity, and normality of errors, we examined residual plots. These plots revealed that:

* Residuals are evenly distributed around zero, indicating no systematic bias in the model's predictions.
* There are no discernible patterns or groupings in the residuals, supporting the assumption of independence.
* The variability of residuals remains consistent across different levels of the independent variable, suggesting homoscedasticity.
* The distribution of residuals closely resembles a normal distribution, fulfilling the normality assumption.

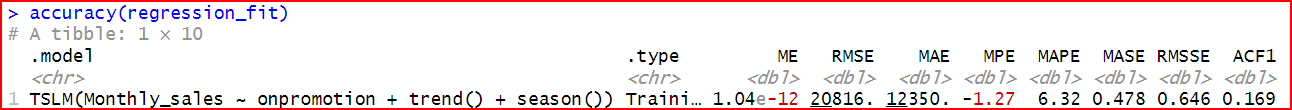
Additionally, the absence of autocorrelation in the residuals further reinforces the assumption of independence.

In conclusion, the analysis of residual plots demonstrates that the model adheres to the key assumptions of regression analysis, bolstering confidence in the validity of its results and forecasts.

The evaluate the regression model and it coeffiecnty we had to generate the following report:

A screenshot of a computer screen

Description automatically generated



Model Evaluation:

While the model exhibits a reasonable fit with an adjusted R-squared of 0.6482, indicating that the independent variables explain 64% of the dependent variable's variation, it's important to consider several other factors when evaluation a model predictive power:

* MAPE of 6.32% suggests a moderate margin of error, with predictions deviating from actual values by an average of 6.32%.
* Statistical significance of only 4 independent variables (intercept, trend, year 8, and year 12 seasonality) at a 0.05 alpha level reveals that many coefficients lack predictive power.

Key takeaways:

* The model provides a general idea of expectations but isn't highly precise.
* Further refinement is necessary to improve accuracy and identify more robust predictors.
* Exploration of additional variables or alternative model specifications might enhance performance.

Recommendations:

* Consider incorporating additional relevant variables or adjusting model features to potentially boost predictive power.
* Clearly communicate the model's limitations and potential for error in forecasting applications to ensure informed decision-making.

Chart of the regression over the actual values:

A graph of a graph showing the growth of a stock market

Description automatically generated with medium confidence

Overall the chart of the model shows that predicted values fit the data well, except for some point where it misses the peaks of the actual data especially in 2017.

**Harmonic Regression:**

Harmonic regression is a specific type of regression that uses trigonometric functions such as sine and cosine to model seasonal patterns in the data. This is particularly useful when the seasonality is a known fixed period with many seasonal coefficients such as daily, or hourly data. The trigonometric functions are represented by the number of Fourier sin and cos pairs denoted as K which is equivalent to including seasonal dummies (Hyndman 2021)

For the purpose of trial and error we used k = 1-6 , and will determine which one is best by AIC and error.

fit\_harmonic <- model(monthly\_stor3\_grocery1,

`K = 1` = ARIMA(log(Monthly\_sales) ~ fourier(K=1) + PDQ(0,0,0)),

`K = 2` = ARIMA(log(Monthly\_sales) ~ fourier(K=2) + PDQ(0,0,0)),

`K = 3` = ARIMA(log(Monthly\_sales) ~ fourier(K=3) + PDQ(0,0,0)),

`K = 4` = ARIMA(log(Monthly\_sales) ~ fourier(K=4) + PDQ(0,0,0)),

`K = 5` = ARIMA(log(Monthly\_sales) ~ fourier(K=5) + PDQ(0,0,0)),

`K = 61` = ARIMA(log(Monthly\_sales) ~ fourier(K=6) + PDQ(1,0,0)),

`K = 6` = ARIMA(log(Monthly\_sales) ~ fourier(K=6) + PDQ(0,0,0)))

A graph of different colored lines

Description automatically generated with medium confidence

> accuracy(fit\_harmonic)

# A tibble: 7 × 10

.model .type ME RMSE MAE MPE MAPE MASE RMSSE ACF1

*<chr>* *<chr>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 K = 1 Training 3586. 31053. 19935. -0.0935 9.51 0.771 0.964 -0.0101

2 K = 2 Training 75.2 26142. 18726. -1.03 8.56 0.724 0.811 0.0184

3 K = 3 Training 1157. 24701. 16509. -0.631 7.97 0.639 0.767 -0.00470

4 K = 4 Training 844. 21709. 12833. -0.538 6.38 0.496 0.674 -0.0644

5 K = 5 Training -721. 21469. 11736. -1.17 5.88 0.454 0.666 -0.167

6 K = 61 Training -735. 21484. 11720. -1.18 5.87 0.453 0.667 -0.169

7 K = 6 Training -719. 21475. 11737. -1.17 5.88 0.454 0.667 -0.168

Model Selection Using AIC:

Understanding AIC:

* AIC (Akaike Information Criterion) balances model fit and complexity.
* Lower AIC values generally indicate better models, as they achieve a good fit without overfitting.

Applying AIC in Harmonic Regression:

* Harmonic regression uses Fourier terms to capture seasonal patterns.
* The number of Fourier terms (K) determines model complexity.
* AIC helps select the optimal K that balances fit and complexity.

Key Points from Your Information:

* AIC was used to select the best K for your harmonic regression model.
* K = 4 yielded the lowest AIC of -57, suggesting it's the optimal choice.
* This implies 4 Fourier terms effectively model the seasonality in data.

**Dynamic regression:**

Dynamic regression combines ARIMA models with regression models. It uses the past values of the series and the past forecast errors to predict future values, while also incorporating additional information from independent variables.

This technique is particularly useful when there are significant external factors influencing the time series that are not captured by ARIMA models alone.

A close-up of numbers

Description automatically generated

Model Structure: ARIMA(1,0,0) with seasonality and trend and onpromotion independent variable.

Yt= 200780.82 + 18.4268Ox1+ 530.5592trend -26734.986season2+8933.749season3 + 5779.737season4+ 4584.592season5-8327.173season6-6783.738season7 -27342.64season8 -10282.07season9+ 259.8373 season10-4189.147season11 + 50444.16season12+ nt

ηt=.6155ηt−1,

εt∼NID(0,516378276).

Here's a breakdown of the model structure and its components:

1. Model Type:

* ARIMA(1,0,0) with seasonality and trend: This indicates a dynamic regression model that includes:
  + Autoregressive component (AR(1)): The current value of the dependent variable (Yt) is partially dependent on the previous value (Yt-1), with a coefficient of 0.6155.
  + Seasonal component: It accounts for recurring patterns within a year, using 11 dummy variables (season2 to season12) to capture monthly variations.
  + Trend component: It represents a linear trend over time, included as a separate variable (trend) in the model.

2. Independent Variable:

* x1: represents the promotional activity, has a coefficient of 18.4268. It suggests that a unit increase in x1 is associated with an estimated increase of 18.4268 units in Yt, holding other variables constant.

3. Error Structure:

* ηt: This represents the model's errors, which are assumed to follow an autoregressive process of order 1 (AR(1)). The coefficient of 0.6155 indicates that a portion of the error in the current period is carried over to the next period.
* εt: These are the independently and identically distributed (i.i.d.) random errors, assumed to have a normal distribution with a mean of 0 and a variance of 516378276.

4. Interpretation of Coefficients:

* Intercept (200780.82): The estimated value of Yt when all independent variables are zero (e.g., no promotion, no trend, and the reference month).
* Trend (530.5592): The estimated increase in Yt for each unit increase in the trend variable, holding other variables constant.
* Seasonal coefficients: Each coefficient represents the estimated difference in Yt for a specific year compared to the reference year 2  is represented as season2, holding other variables constant.

1. Regression Assumptions:

**A graph with a line graph

Description automatically generated**

A screenshot of a graph

Description automatically generated

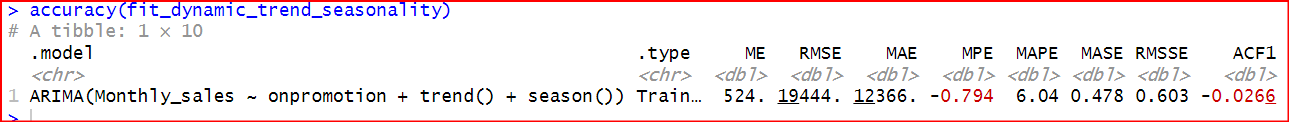
A graph of a graph of a number of days

Description automatically generated with medium confidence

**The observed positive linear relationship within the data supports the linearity assumption of regression analysis. Furthermore, the visual inspection of the residual plot, resembling white noise, suggests that the assumptions of independence and normality of errors are also likely met. This indicates that the errors are randomly distributed, uncorrelated with each other, and follow a normal distribution.** **Adherence to these assumptions is crucial for ensuring the validity of the regression model's results and the reliability of its inferences.**

1. **Evaluating the models predictive power evaluation:**

AIC=1297.43 AICc=1311.38 BIC=1329.83

****

The dynamic ARIMA regression appears to be the most accurate model among all the models I developed so far in this report. I based this on the following key metrics:

* MAPE (Mean Absolute Percentage Error): At 6.04%, it boasts the lowest MAPE among all models so far, indicating that on average, its predictions deviate from actual values by only 6.04%. This image further illustrates how well the dynamic regression model fits the actual values :

A graph with red and black lines

Description automatically generated

* RMSE (Root Mean Squared Error): While the RMSE of 19,444 might seem high, it's essential to consider the scale of your dependent variable. In the context of your specific data, this RMSE could still represent reasonably accurate forecasts. For example the mean monthly sales for store 3 grocery department is. 231615.5
* AIC (Akaike Information Criterion): The AIC value of 1297.43 is the lowest among all the candidate models, suggesting a good balance between model fit and complexity.

**Prophit Regression**

Prophet regression, also known as Facebook Prophet, is a forecasting technique designed specifically for time series data. It's an open-source, automated forecasting procedure developed by Facebook's Core Data Science team.

Model Breakdown:

Prophet falls under the category of additive models, meaning it forecasts future values by summing up several components:

* + Trend: This captures the overall direction of the time series, whether it's increasing, decreasing, or remaining stable. It's modeled using piecewise linear regression, allowing for flexibility in fitting the trend.
  + Seasonality: This component accounts for recurring patterns in the data, like weekly, monthly, or yearly trends. It uses Fourier series to model different seasonal patterns.
  + Holidays: Prophet allows explicitly including the impact of known holidays on your data.
  + Noise: This component captures any random fluctuations in the data not explained by the other factors.

Strengths:

* Easy to use: Prophet is designed for ease of use and requires minimal data wrangling or parameter tuning.
* Fast and efficient: It's well-optimized for speed and scalability.
* Handles diverse data: It's robust to missing data, shifts in trends, and outliers, making it suitable for real-world datasets.
* Visualizations: Prophet provides built-in visualization tools to analyze and interpret the forecasts.

Limitations:

* Univariate: Prophet only forecasts a single target variable, not multiple variables simultaneously.
* Black box nature: While flexible, it can be challenging to understand the inner workings of the model and interpret its coefficients.Especial since there is no report function in prophet.
* Better for strong seasonality: It performs best with time series exhibiting strong seasonal patterns and requires sufficient historical data to learn these patterns effectively.

Overall, Prophet regression is a powerful and user-friendly tool for time series forecasting, particularly when dealing with data exhibiting trends, seasonality, and holiday effects. Its ease of use and fast performance make it a popular choice for diverse forecasting needs.

**A graph of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated**

accuracy(fit\_profit\_month)

# A tibble: 1 × 10

.model .type ME RMSE MAE MPE MAPE MASE RMSSE ACF1

*<chr>* *<chr>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 profit Training 0.177 12709. 8032. -0.585 4.06 0.311 0.394 0.222

Based solely on error metrics the prophet model seem to perform the best and has the lowest error.

**Conclusion and Key Findings:**

* Profit Model Outperforms: The profit model demonstrates the most accurate forecasting capabilities for grocery sales at store 3, with a significantly lower RMSE of 12,709 and a MAPE of 4.06%.
* Dynamic Regression Model as Runner-Up: While unable to calculate AIC for the profit model, the dynamic regression model emerges as the second-best option, exhibiting a MAPE of 6.04% and RMSE of 19,444.
* Context for Evaluation: The average monthly sales for store 3 grocery department are $231,615 , providing a benchmark for assessing model performance.

Limitations and Considerations for time series forcasting using Fable Library:

* Restricted Data: The techniques used relied primarily on past time series data, omitting potentially influential variables such as daily oil prices and regional holiday information due to my time constraints.
* Forecasting as a Guide, Not a Guarantee: It's crucial to acknowledge that forecasting is not an infallible tool. It should be employed as a reference, complemented by domain expertise and judgmental forecasting to capture a more comprehensive picture.

Recommendations:

* Incorporate Additional Variables: Explore the inclusion of daily oil prices, regional holiday information, and other pertinent factors to potentially enhance model accuracy.
* Prioritize Domain Knowledge: Integrate insights from individuals with deep understanding of the grocery industry and store-specific dynamics to enrich forecasting efforts.
* Combine Forecasting Methodologies: Strategically blend statistical models with judgmental forecasting techniques to leverage the strengths of both approaches and mitigate potential biases.

Conclusion:

While the profit model exhibits promising predictive power for grocery sales at store 3, it's essential to recognize the limitations of time series forecasting and the importance of complementing it with both domain expertise and informed judgment to achieve robust and reliable business decisions. For example, no one could forecast that covid 19 was going to turn our world upside down, and it is especial hard to forecast cycle, and recession. Overall, I learned a valuable skill and this class boosted my confidence that I can teach myself extremely complex mathematical modeling, and overcome many hurtles to reach my goals. Thank you for taking the time to help me and I hope you enjoyed my report.,