**Forecasting Electricity Generation in Ireland: A Time Series Analysis**

**&**

**Identifying Frequent Itemset and the associated Association rules: Retail store**

An exploratory study using EDA, Statistical Analysis, and Machine Learning

**Author:**

Selvi Ganesh Kumar

**Institution:**

CCT College

Dublin 2

Ireland

**Course:**

Higher Diploma in Science in Data Analytics

**Supervisor:**

Mr. Sam Weiss

**Date of Submission:**

12th of April 2025

**Version:**

Final Report

**Contact:**  
Email: [selvi.digitalsolutions@gmail.com]  
GitHub: github.com/research-selvi-datascience

*© 2025 Selvi Ganesh Kumar. All rights reserved.*

**Forecasting Electricity Generation in Ireland: A Time Series Analysis**

**Table of Contents:**

|  |  |
| --- | --- |
| 1. **Introduction (Total electricity generation and Wind energy Forecasting)** | **3** |
| 1. Methodology | **3** |
| 1. Feature Engineering: Total electricity generation | 3 |
| 1. Time Series Decomposition | 3 |
| 1. Model Selection and Evaluation | 3 |
| 1. Results & Model Evaluation | 4 |
| 1. Performance Metrics Comparison | 6 |
| 1. **Conclusion: Total electricity generation - Forecasting** | **8** |
| 1. **Conclusion: Wind Energy Forecasting** | **9** |
| **Market Basket Analysis** | 9 |
| 1. Introduction | 9 |
| 1. Preprocessing | 10 |
| 1. Experiments | 10 |
| 1. Experiment 1: Apriori (Minimum Support 0.01) | 10 |
| 1. Experiment 2: Apriori (Minimum Support 0.005) | 11 |
| 1. Experiment 3: Apriori (Minimum Support 0.0005) | 11 |
| 1. Experiment 4: FP-Growth (Minimum Support 0.01) | 12 |
| 1. Experiment 5: FP-Growth (Minimum Support 0.005) | 12 |
| 1. Experiment 6: FP-Growth (Minimum Support 0.0005) | 13 |
| 1. Findings | 14 |
| 1. Conclusion | 14 |
| 1. **Appendix: Table of Figures** | **14** |
| 1. **References** | **15** |

**Introduction - Total Electricity Generation and Wind Energy Forecasting**

Electricity generation is crucial to a nation’s energy strategy, impacting economic growth, sustainability, and environmental policies. This report analyzes Ireland’s total electricity generation from January 2010 to December 2024, across sources such as coal, renewables, hydro, natural gas, oil, peat, solar, biomass, and wind.

Additionally, it examines wind energy for the same period. Using time series analysis techniques like ARIMA and SARIMA, the study identifies trends, seasonality, and forecasts future demand. Model performance was evaluated using metrics such as R², AIC, RMSE, MAE, and MAPE, with the best models chosen for their accuracy and interpretability.

**Methodology**

**Total Electricity Generation - Feature Engineering:**

The total electricity generated was calculated for each month of every year across all energy sources to observe long-term patterns and seasonal fluctuations.

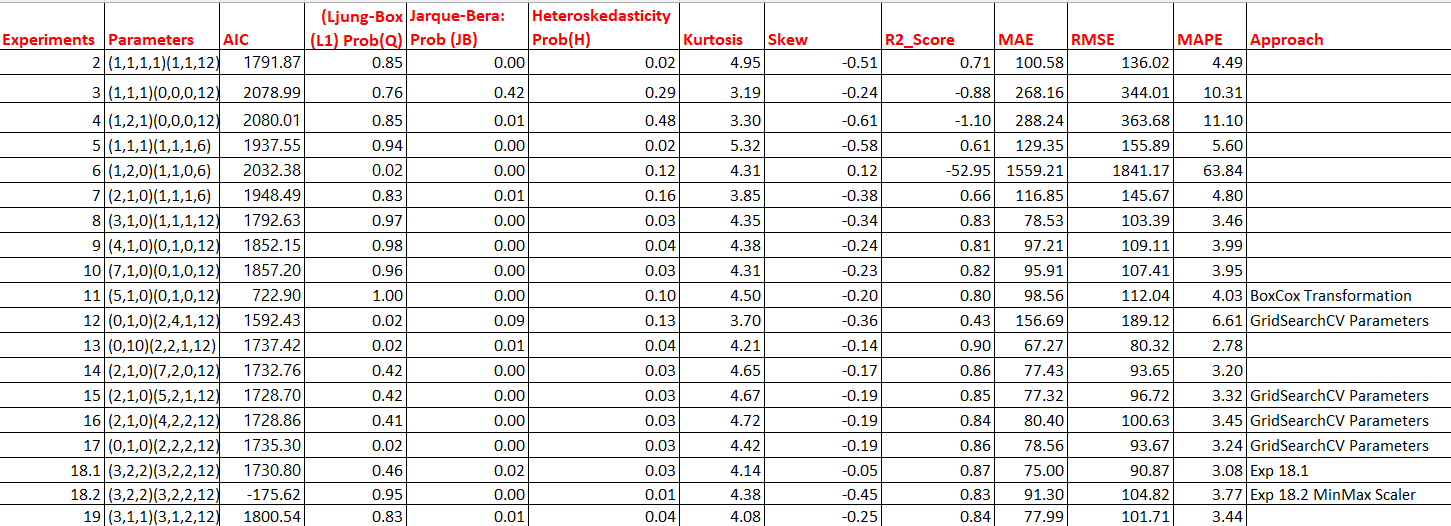
**Time Series Decomposition**

A seasonal decomposition plot showed that electricity generation in Ireland has steadily increased since 2016, with a notable surge from 2019 onwards. The peak occurred between 2022 and 2023, followed by stabilization. Seasonal patterns were consistent throughout the years, with residual analysis revealing occasional deviations, especially in early 2020, 2021, and 2024.

**Model Selection and Evaluation – Total electricity Generation**

To develop robust forecasting models, 20 different experiments were conducted using various combinations of ARIMA/SARIMA parameters. The initial selection of parameters was based on Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots, followed by hyperparameter tuning via GridSearchCV. The following statistical tests were conducted to assess model validity:

* **p-values:** Ensured statistical significance of parameters.
* **Ljung-Box Test:** Checked for residual autocorrelation.
* **Jarque-Bera Normality Test:** Evaluated normality of residuals.
* **Heteroskedasticity Test:** Verified variance stability.
* **Kurtosis Analysis:** Assessed residual distribution shape and tail behaviour.
* **AIC Value:** For better model fit.
* **Performance metrics:** R2\_Scores and Errors (MAPE, MAE, RMSE)

****

***Figure 1: 20 Experiments and approaches with performance scores for Total\_Generation***

**Results & Model Evaluation – Total Generation**

**Seasonal Parameter Effects**

Experiments with varying seasonal parameters revealed that:

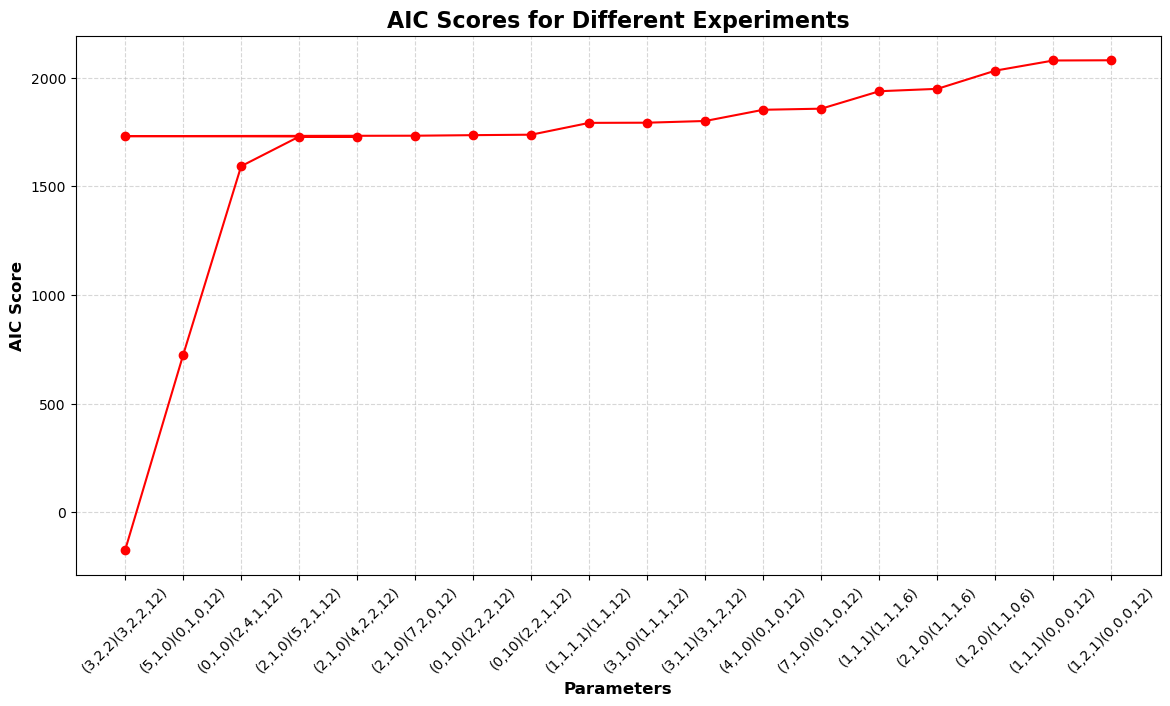
* High differencing values (D > 3) led to increased prediction errors.
* D = 2 provided the best trade-off between accuracy and stability.
* Seasonal P and Q as ‘0’ or ‘1’ resulted in significant performance degradation.

**AIC Scores:** The top three experiments that obtained low AIC scores were:

(5,1,0)(0,1,0,12) with 722.899;

(0,1,0)(2,4,1,12) with 1592.429;

(2,1,0)(5,2,1,12) with 1728.700, though 6-7 models were close to each other with their AIC score.



***Figure 2: AIC scores of all the Model experiments***

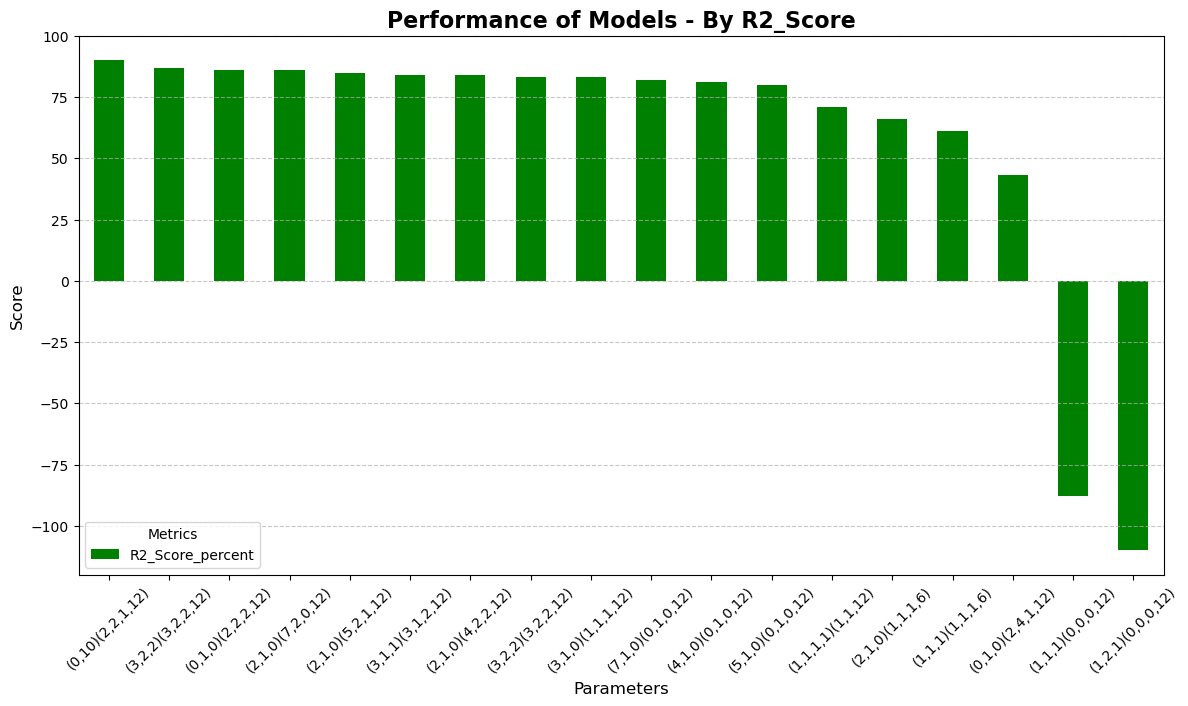
**R2\_Scores:**The top three experiments, that achieved high score was

(0,1,0)(2,2,1,12) with 0.90,

(3,2,2)(3,2,2,12) with 0.87,

(0,1,0)(2,2,212) with 0.86 and

(2,1,0)(7,2,0,12) with 0.86



***Figure 3: R2\_Scores of all parameters***

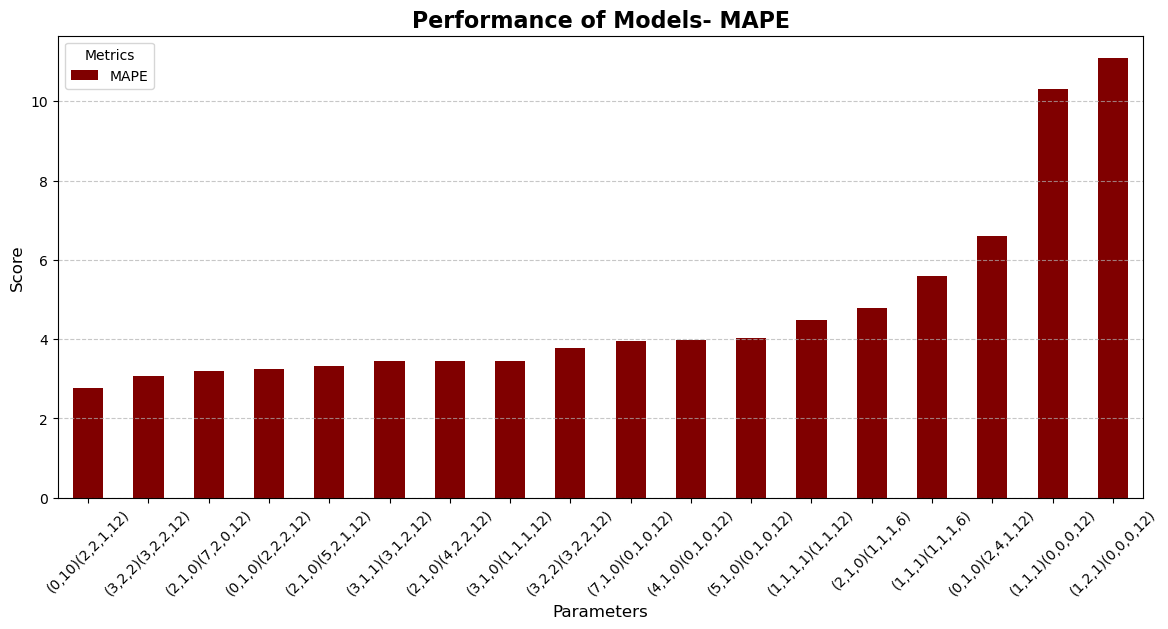
**MAPE Scores:**Thetop four models, that achieved high score was:

(0,1,0)(2,2,1,12) with 2.78%,

(3,2,2)(3,2,2,12) with 3.08%,

(2,1,0)(7,2,0,12) with 3.20% and

(0,1,0)(2,2,2,12) with 3.24%.



***Figure 4: MAPE Scores of all parameters***

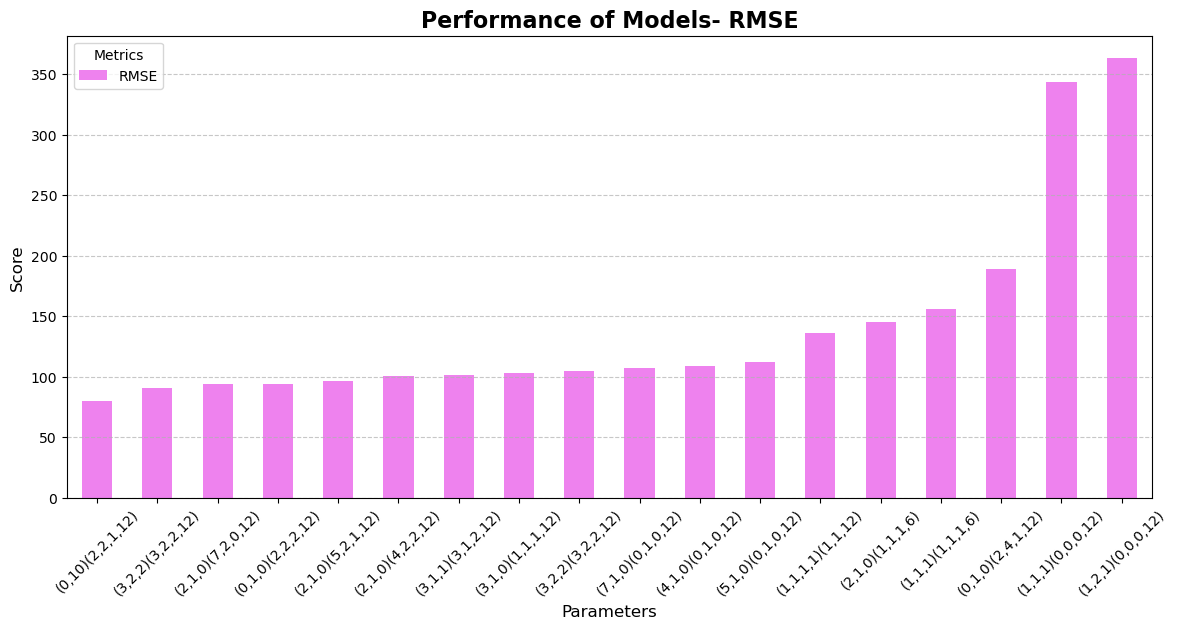
**RMSE Scores:**The top four experiments, that achieved high score was

(0,1,0)(2,2,1,12) with 80.32,

(3,2,2)(3,2,2,12) with 90.87,

(2,1,0)(7,2,0,12) with 93.65 and

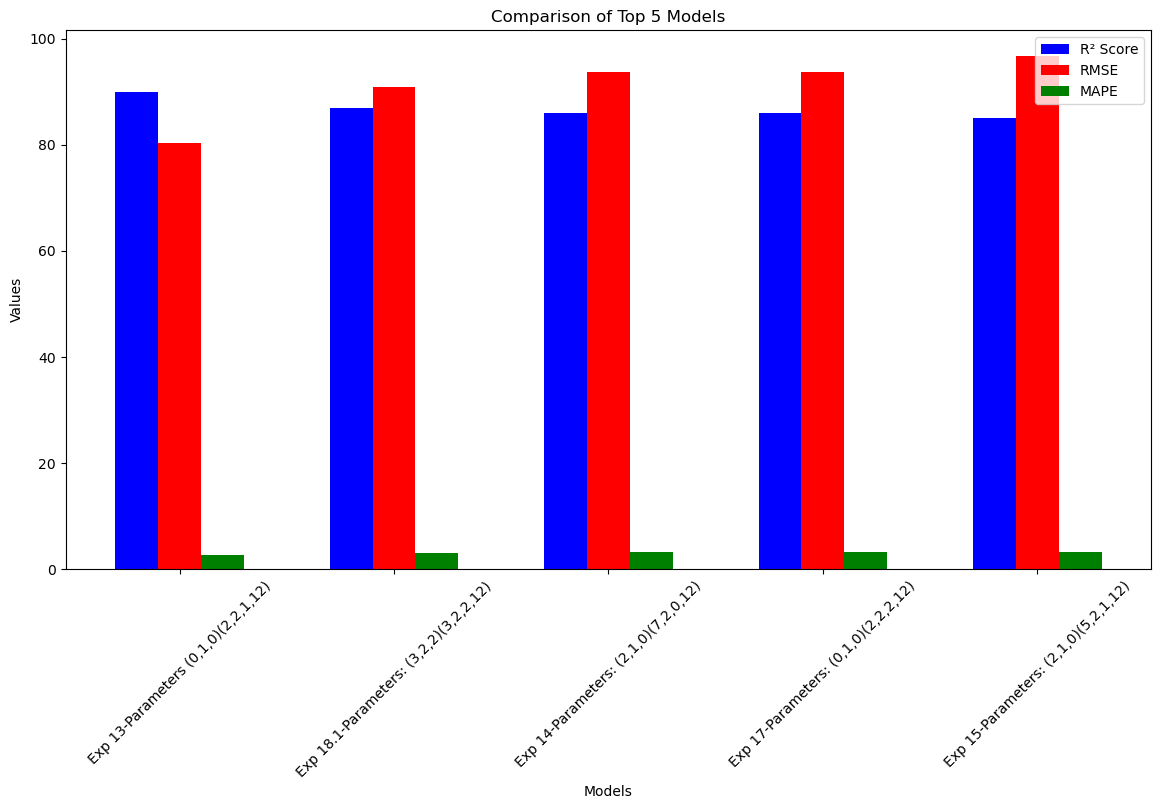
(0,1,0)(2,2,2,12) with 93.67



***Figure 5: RMSE Scores of all parameters***

**Performance Metrics Comparison**

The top five models were selected based on the above key performance indicators:



***Figure 6: Performance of Top 5 Models – Total\_Generation***

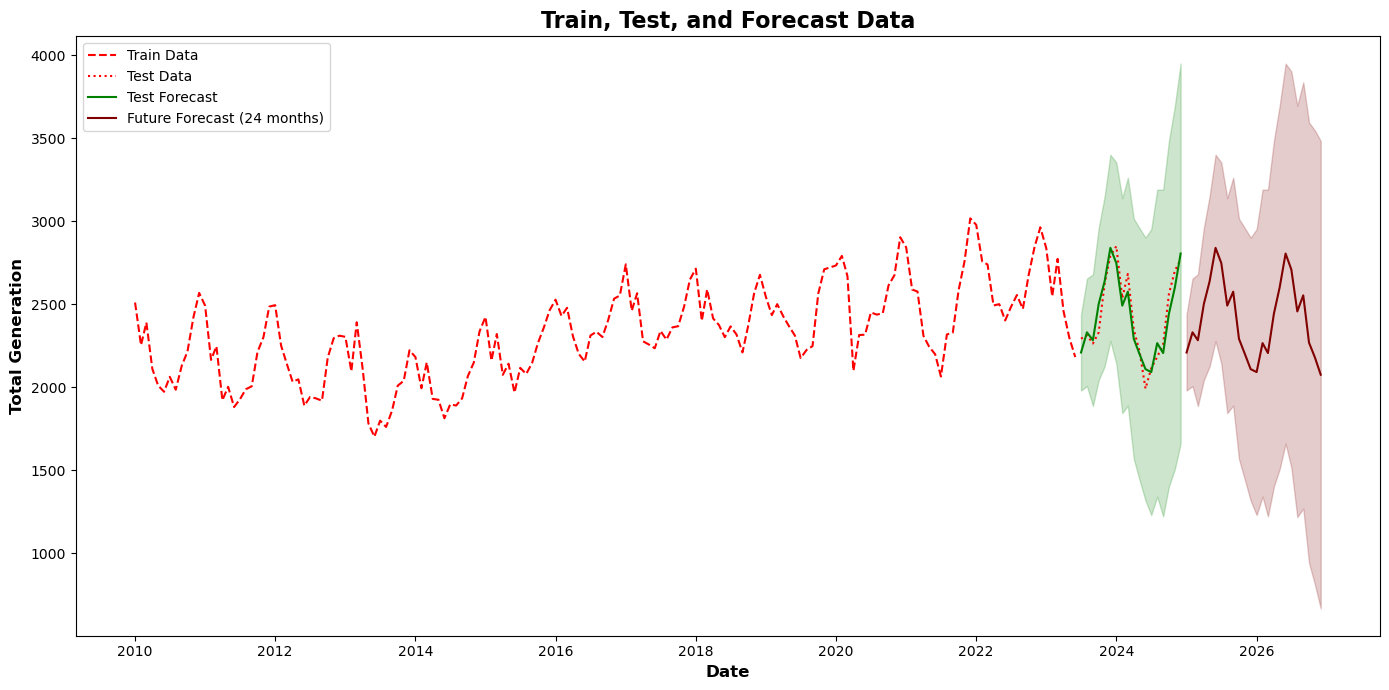
**Experiment 13 (Best Model): (0,1,0)(2,2,1,12)**

**R² Score:** 90% (Highest)

**RMSE**: 80.32 (Lowest)

**MAPE:** 2.78% (Lowest)

**AIC:** 1737.415 (Higher, but acceptable given the strong fit and accuracy)



*Figure 7: Train, Test, Forecast plot of the Experiment 13 – (0,1,0) (2,2,1,12)*

**Residual Analysis & Model Diagnostics**

Key findings from the best-performing model (Experiment 13) included:

* **Autoregressive Coefficients:** Significant seasonal lags at 12 and 24 months.
* **Residual Tests:** Ljung-Box (p = 0.02) showed some autocorrelation; Jarque-Bera (p = 0.01) indicated non-normality; heteroskedasticity (p = 0.04) suggested slight variance fluctuations and mild kurtosis.

**Conclusion:** This model provides the most accurate predictions and the best fit.

|  |  |
| --- | --- |
| **The next best four models, its parameters and the performance scores** | |
| * **Experiment 18.1: (3,2,2)(3,2,2,12)**   + **R² Score:** 87%   + **RMSE:** 90.87   + **MAPE:** 3.08%   + **AIC:** 1730.796 (Relatively low, in comparison with the rest of the models)   + **Conclusion***:* This model is also highly accurate and generalizable. | * **Experiment 14: (2,1,0)(7,2,0,12)**   + **R² Score:** 86%   + **RMSE:** 93.65   + **MAPE**: 3.20%   + **AIC:** 1732.761 (Relatively low)   + **Conclusion:** Another strong candidate with balanced accuracy and fit. |
| * **Experiment 17: (0,1,0)(2,2,2,12)**   + **R² Score:** 86%   + **RMSE:** 93.67   + **MAPE:** 3.24%   + **AIC:** 1735.299   + **Conclusion:** Despite slightly higher RMSE/MAPE, indicating lower complexity. | * **Experiment 15: (2,1,0)(5,2,1,12)**   + **R² Score:** 85%   + **RMSE**: 96.72   + **MAPE:** 3.32%   + **AIC:** 1728.700 (Relatively low, a best fit with higher R2 Score and low errors) |

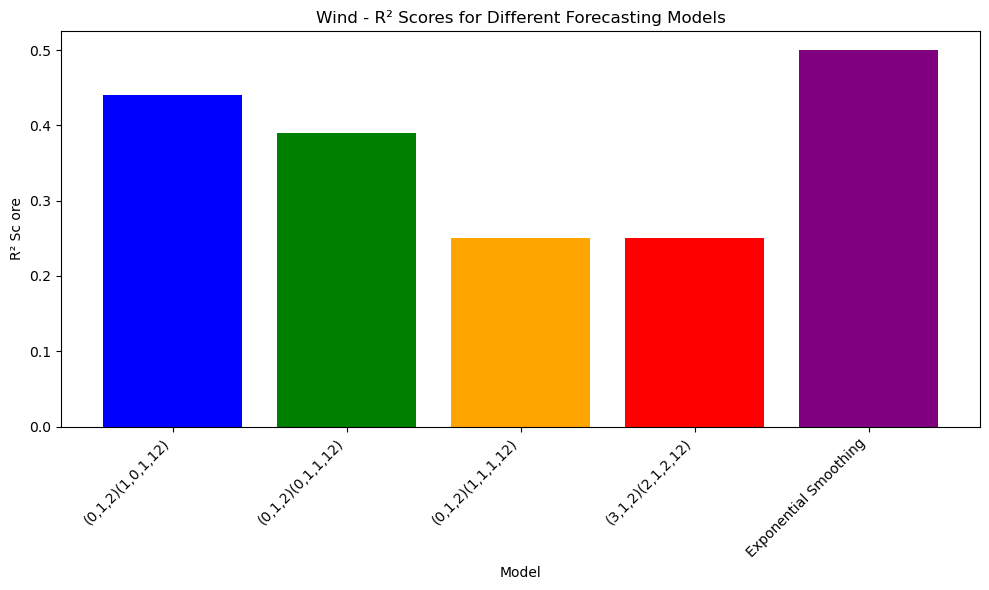
**Conclusion: Total electricity generation - Forecasting:**

This study successfully developed time series forecasting models for total electricity generation in Ireland. The best-performing model (Experiment 13) provided a high R² (90%) and low MAPE (2.78%), making it suitable for accurate predictions. Seasonal factors played a crucial role in model performance, emphasizing the need to incorporate them into forecasting models.

While some residual issues (autocorrelation and variance fluctuations) remain, the selected model provides reliable forecasts for future electricity demand.

**Conclusion : Wind Energy Forecasting:**

Wind energy forecasting was conducted as part of Ireland's electricity generation analysis, driven by significant growth, increasing demand, and the global shift towards renewables.



**Figure 8: R2\_Scores for Different Wind Forecasting models**

Exponential Smoothing achieved a higher R² score of 50%, capturing half of the variance, while SARIMA had an AIC of 2085.117 and an R² score of 44%. Both models showed normal residuals and no significant autocorrelation, but exhibited heteroscedasticity and kurtosis, indicating non-constant variance and medium-tailed distributions.

Future improvements could include refining seasonal components, testing deep learning approaches, or integrating external factors (economic indicators, policy changes) for enhanced forecasting accuracy.

**Market Basket Analysis:**

**Introduction**

Market Basket Analysis identifies relationships between frequently purchased items using Apriori (iterative frequent itemset generation) and FP-Growth (compact tree-based approach). This study analyzes grocery store transactions to uncover strong product associations. Key Metrics:

* Support: Itemset frequency
* Confidence: Rule reliability
* Lift: Rule strength
* Conviction: Predictive power & dependency

**Preprocessing**

1. **Data Transformation:** The dataset was structured into a list of transactions representing purchased items.
2. **One-Hot Encoding:** A binary matrix was created where rows represent transactions, columns represent unique items, and each cell indicates item purchase.

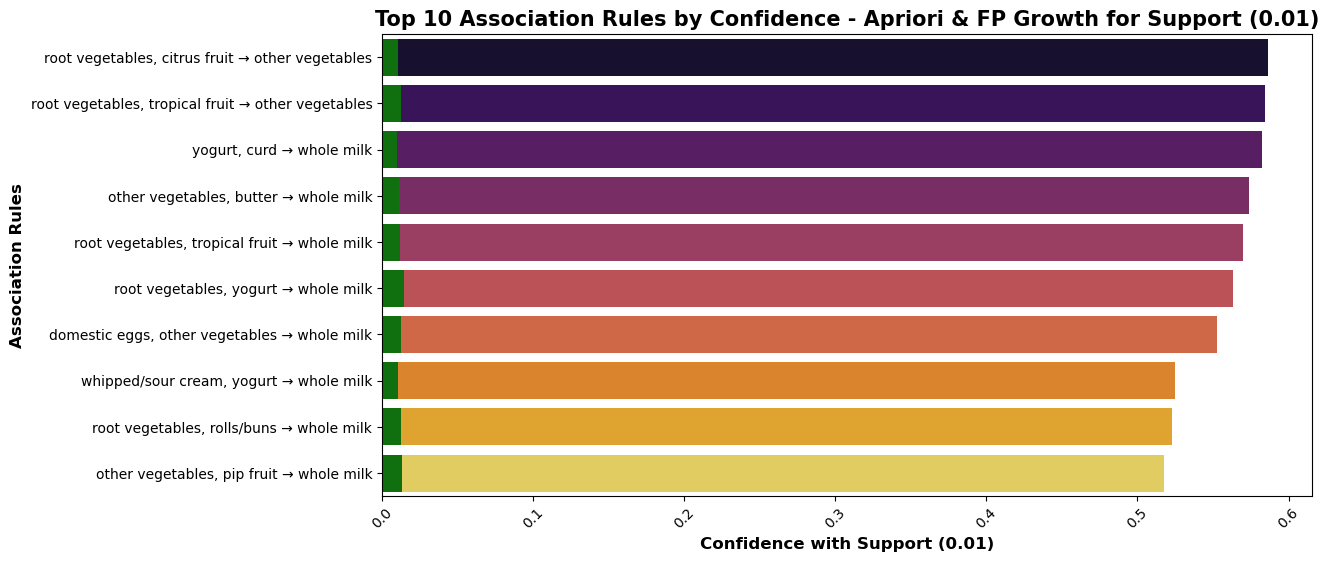
**Experiments**

* **Frequent Itemset Mining:** Both Apriori and FP-Growth algorithms were applied with different minimum support levels (0.01, 0.005, and 0.0005) to extract frequent itemsets.
* **Association Rule Generation:** Rules were generated by varying confidence thresholds (0.5, 0.3, 0.1).
* **Analysis:** The top 10–20 association rules were evaluated using Lift, Confidence, and Conviction, revealing different patterns based on varying support and confidence levels.

**Experiment 1: Apriori (Minimum Support 0.01)**

**Confidence 0.5:**

High-lift values indicate strong product relationships for bundling and promotions. The top 10 associations highlights the combination of fresh produce and whole milk.



**Figure 9: Top 10 Association Rules by Confidence for support (0.01)**

**Confidence 0.3:**

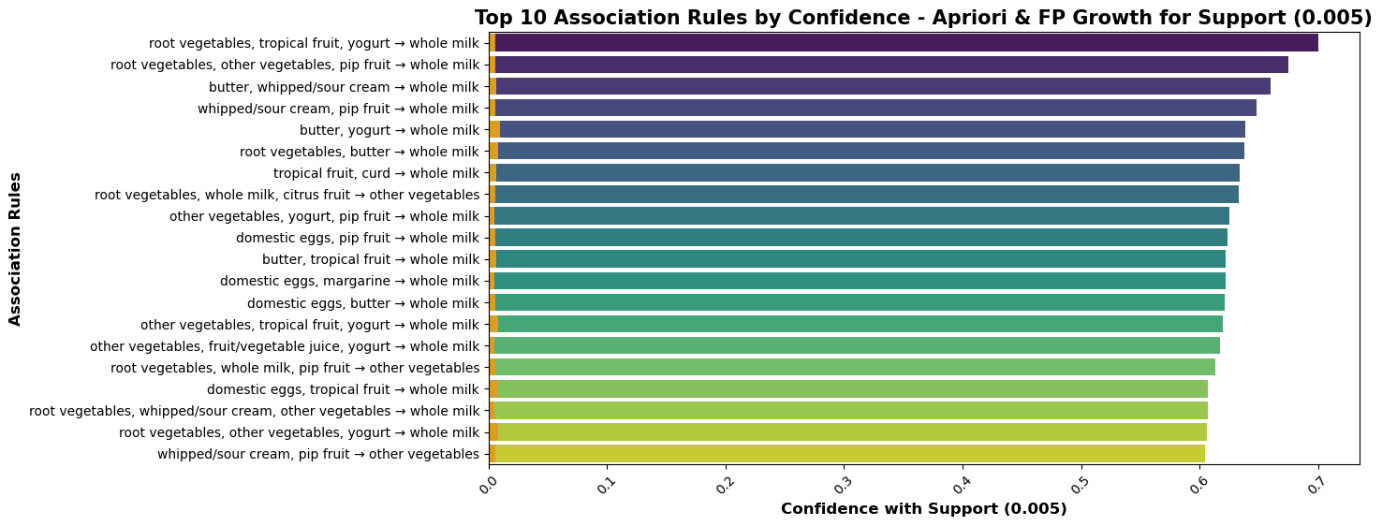
* + Beef’s repeated associations suggest cross-selling potential. High-lift associations (e.g., beef with vegetables, baking powder with milk) should be prioritized for cross-promotion and store placement.

**Confidence 0.1:**

* + More associations appear, revealing lower-confidence but still insightful relationships.

**Experiment 2: Apriori (Minimum Support 0.005)**

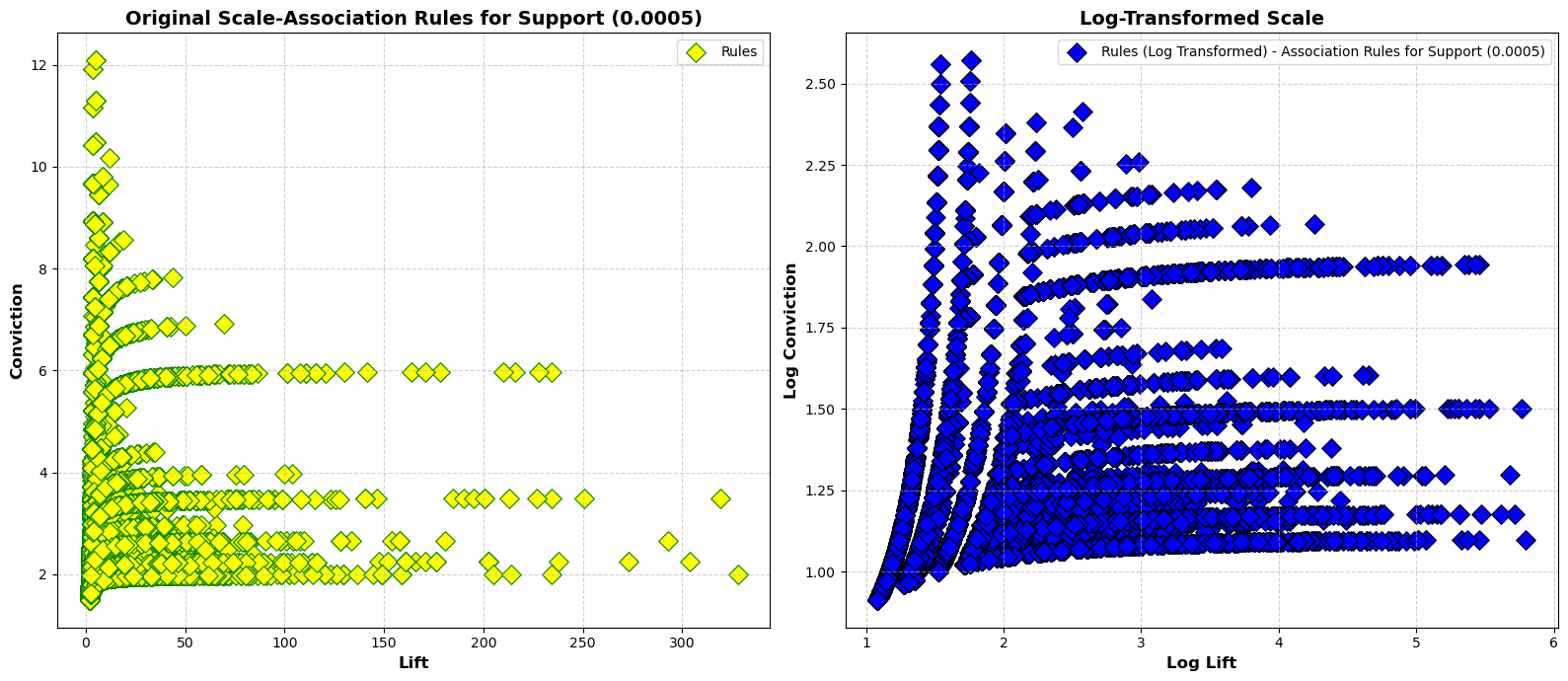
Lowering minimum support to 0.005 increased frequent itemsets to 1001. The same confidence thresholds (0.5, 0.3, 0.1) were applied to uncover significant product relationships. Similar trends emerged, confirming the robustness of associations.



* **Figure 10: Top 10 Association Rules by Confidence for support (0.005)**

**Experiment 3: Apriori (Minimum Support 0.0005)**

Reducing minimum support further to 0.0005 increased frequent itemsets to 49,579, capturing rare item associations and larger item combinations with high lift, conviction and confidence levels.



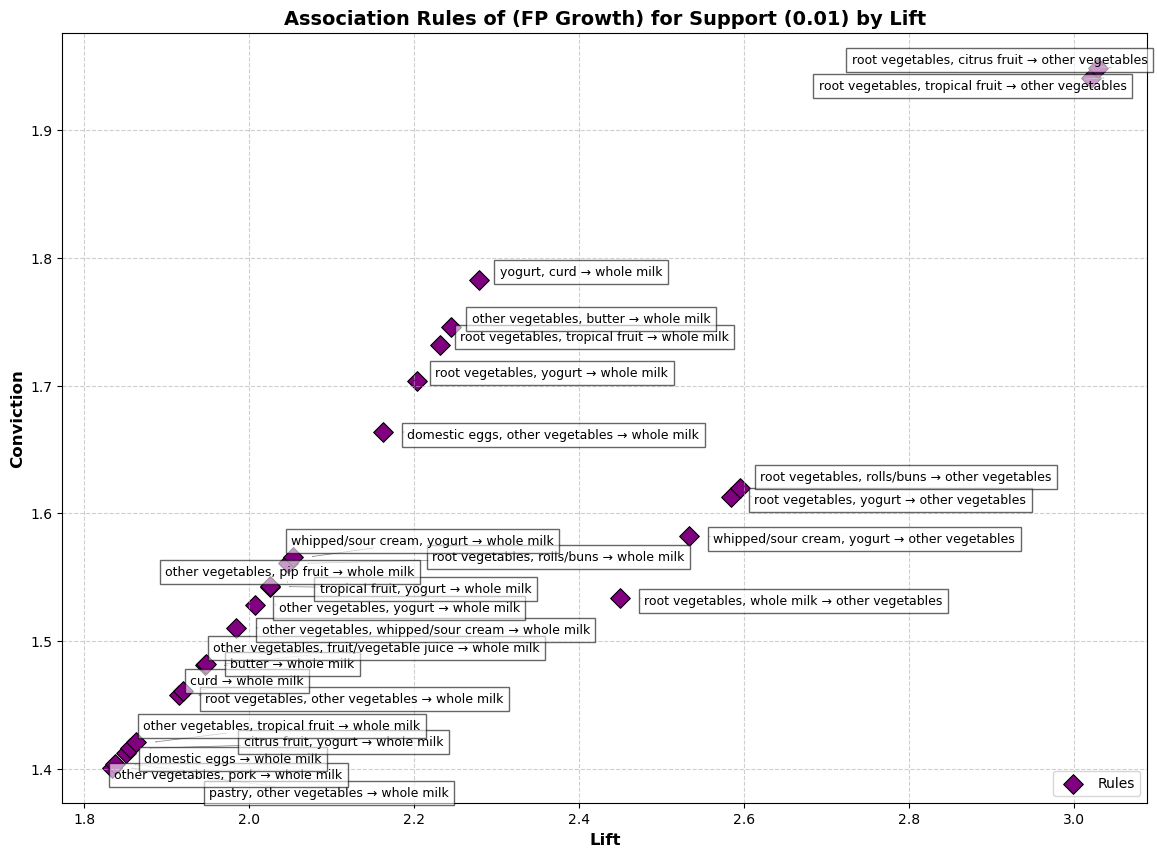
* **Figure 11: Association Rules by Confidence for support (0.0005)**

**Key Findings:**

* + Rare items like baby cosmetics and artificial sweeteners emerged.
  + Larger itemsets (3-4 items) appeared, such as (whipped/sour cream, margarine, rolls/buns, root vegetables).

**Experiment 4: FP-Growth (Minimum Support 0.01)**

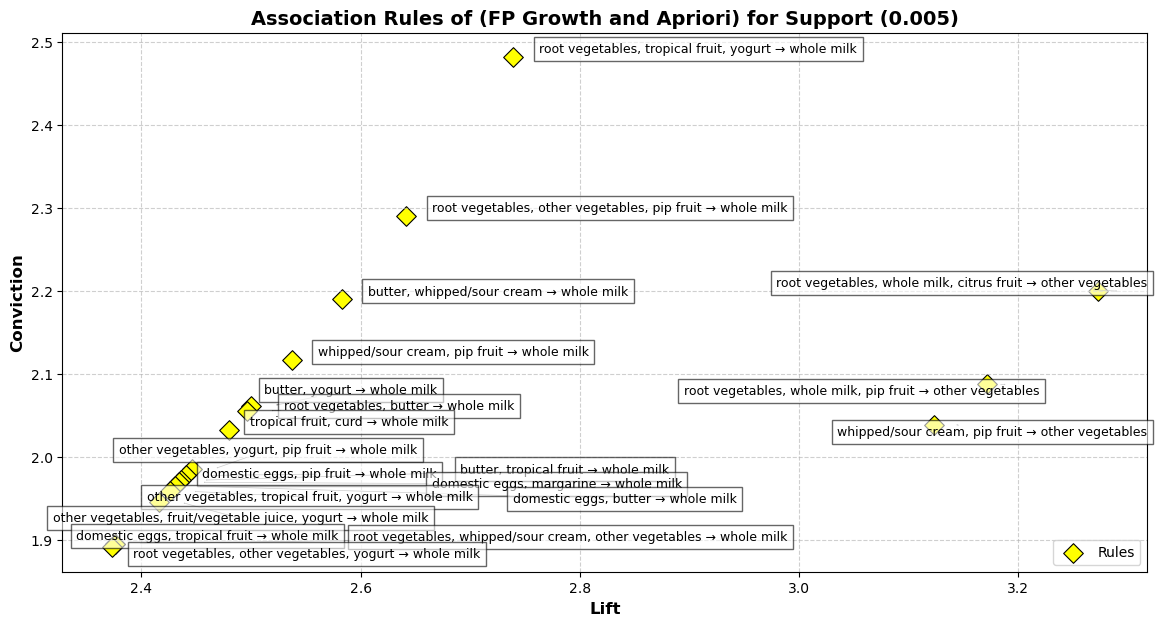
The analysis revealed same results as Apriori. So, analysed with Confidence ≥ 0.46.The whole milk and vegetables ranked high as consequents. So, pairing whole milk with fresh produce and breakfast essentials and positioning dairy near fruits, vegetables, and pastries would benefit.



**Figure 12: Association Rules by Lift for support (0.01)**

**Experiment 5: FP-Growth (Minimum Support 0.005)**

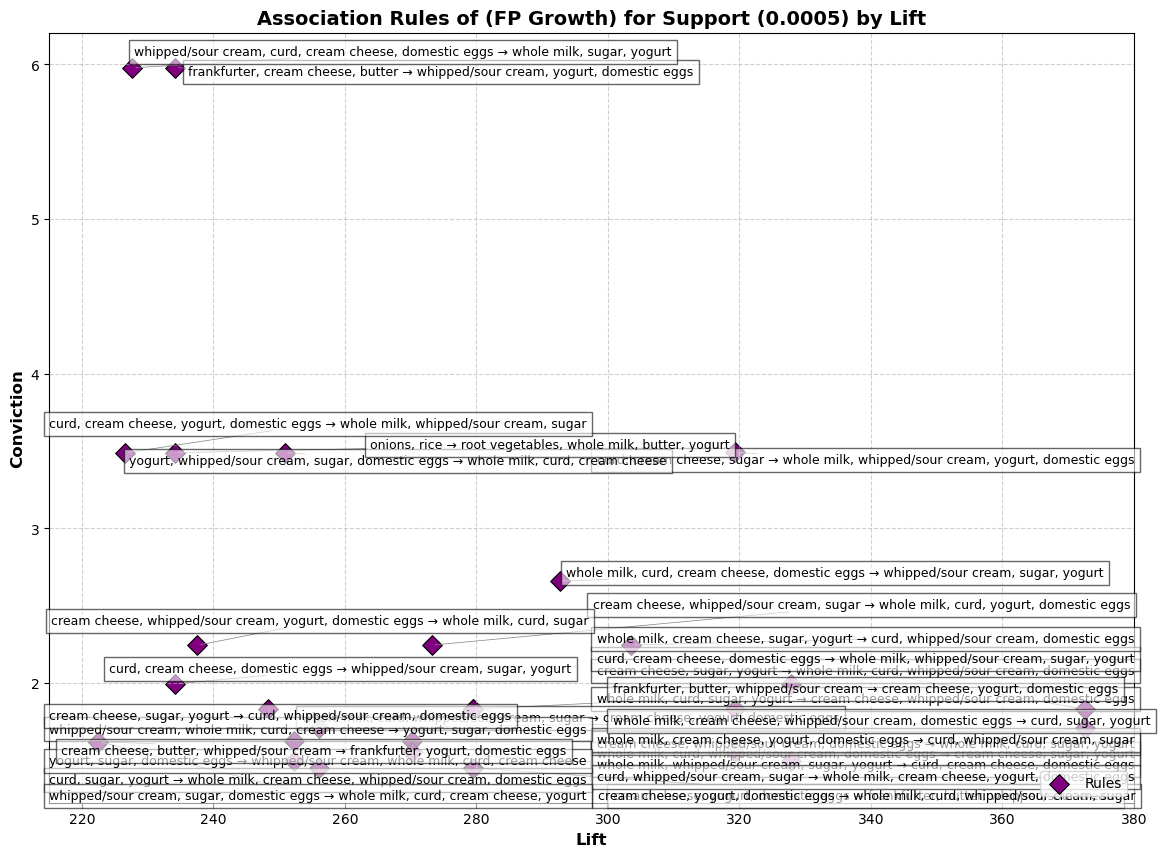
Revealed same results like Apriori, which were visualized using scatter plots to compare association strengths.



**Figure 13: Association Rules by Lift vs Conviction for support (0.005)**

**Experiment 6: FP-Growth (Minimum Support 0.0005)**

Certain associations have resulted in Lift more than 350 and conviction rate of upto 12. This reveals serious niche associations (eg. whole milk, yogurt and sugar together) for specific targeted approaches.



**Figure 14: Top Association Rules by Lift vs Conviction for support (0.0005)**

**Findings:**

* **Frequent Product Associations:** Whole milk frequently appeared as a consequent, indicating high demand.
* **Lowering Support Thresholds:** Detected more associations, including rare item combinations.
* **Confidence, Lift & Conviction Metrics:** Highlighted strong purchasing patterns for product placement and cross-selling.
* **Algorithm Performance:** FP-Growth was computationally efficient, while Apriori required more iterations.
* **Business Applications:** Supports product bundling, cross-selling, and store layout optimization to enhance sales, marketing, and customer experience.

**Conclusion:**

* Both algorithms identified frequent itemsets and association rules consistently across support levels (0.01, 0.005, 0.0005), with nearly identical metrics (support, confidence, lift, leverage, conviction). While both used the same thresholds, FP-Growth was significantly faster, making it more efficient for large datasets, whereas Apriori was more computationally intensive but provided similar insights.

**Appendix**

**Table of Figures**

|  |  |
| --- | --- |
| 1. Figure 1: Experiments and approaches with performance scores | 4 |
| 1. Figure 2 : AIC Scores of all parameters | 4 |
| 1. Figure 3: R2\_Scores Scores of all parameters | 5 |
| 1. Figure 4: MAPE Scores of all parameters | 6 |
| 1. Figure 5: RMSE Scores of all parameters | 6 |
| 1. Figure 6: Performance of Top 5 Models – Parameters | 7 |
| 1. Figure 7: Train, Test, Forecast plot of the Experiment 13 – (0,1,0) (2,2,1,12) | 7 |
| 1. Figure 8: R2\_Scores for Different Wind Forecasting models | 9 |
| 1. Figure 9: Top 10 Association Rules by Confidence for support (0.01) | 10 |
| 1. Figure 10: Top 10 Association Rules by Confidence for support (0.005) | 11 |
| 1. Figure 11: Association Rules by Confidence for support (0.0005) | 11 |
| 1. Figure 12: Association Rules by Lift for support (0.01) | 12 |
| 1. Figure 13: Association Rules by Lift vs Conviction for support (0.005) | 13 |
| 1. Figure 14: Top Association Rules by Lift vs Conviction for support (0.0005) | 13 |

**References:**

1. Tutorial\_MBA\_Apriori.ipynb, Mr. Sam Weiss

2. Tutorial\_MBA\_Apriori - In class.ipynb, Mr. Sam Weiss

3. MBA-Actual-Class.ipynb, Mr. Sam Weiss

4. Tutorial 9 - MBA.ipynb, Dr. Mohammed Iqbal

5. Apriori - Store dataset.ipynb, Ms. Marina Iantorno

6. PDA - Market Basket Analysis.ipynb, Ms. Marina Iantorno

7. ARMA - 1.ipynb, Mr. Sam Weiss

8. ARMA - 2 and ARIMA.ipynb, Mr. Sam Weiss

9. SARIMA Full.ipynb, Mr. Sam Weiss

10. SARIMA.ipynb, Mr. Sam Weiss

11. TSA-SARIMA-US Air Traffic Part 1.ipynb, Mr. Sam Weiss

12. TSA-SARIMA-US Air Traffic Part 2.ipynb, Mr. Sam Weiss