Méabh Hartney

**ABSTRACT**

This research explores the accuracy of forecasting retail alcohol sales and the factors that play a role in the process. I used three FRED datasets, including US Retail Sales, Gross Domestic Product, and Consumer Price Index. I split the data into training and testing and evaluated four simple forecasting models, followed by multivariate regression models, ARIMA, and exponential smoothing models. I found that the regression model, using season, GDP, and CPI as independent variables, is the best model for forecasting alcohol sales. The study highlights the impact of the COVID-19 pandemic on retail alcohol sales and the economy in general.

Keywords: Retail alcohol sales, forecasting, RStudio, GDP, CPI, COVID-19 pandemic, FRED datasets, multivariate regression models, ARIMA, exponential smoothing models and seasonality.

1. **INTRODUCTION**

I was interested in determining whether retail alcohol sales could be forecast accurately. I wanted to establish what, if any, factors were important, when attempting to forecast. I was also very interested in examining what the impact of the COVID-19 pandemic was on retail alcohol sales and the effect that this extraordinary event would have on the models.

1. **DATA**

Data was sourced from the US Federal Reserve Economic Data (FRED) St. Louis Fed. I used three FRED datasets, with quarterly non-seasonal data from 1992-2022.

Datasets:

US Retail Sales: Beer, Wine and Liquor Stores (Figure 2.1)

US Gross Domestic Product (Figure 2.3)

Consumer Price Index for All Urban Consumers: Alcoholic Beverages in U.S City Average. (Figure 2.4)

Using the Guerrero method, I determined that a log transformation of the sales data was appropriate. (Figure 2.2)

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Fig 2.1 – Quarterly sales from 1992-2022 Fig 2.2 - Log transformation

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Fig 2.4 – Quarterly CPI for alcoholic beverages from 1992-2022

Fig 2.3 – Quarterly GDP from 1992-2022

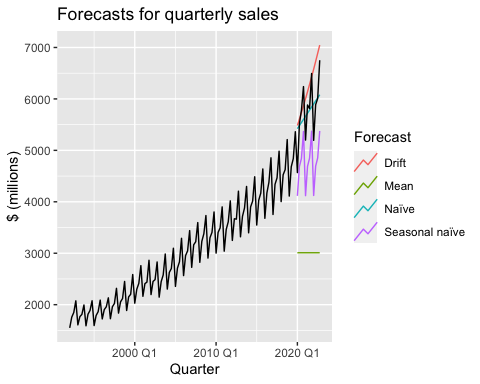
TEST AND TRAINING

I split the data into training – 1992 until 2019 - and tested on the post-2020 data. As the effect of Covid-19 pandemic was first felt in the US from March 2020, I used the sales data from Q1 2020 onwards as out-of-sample testing data to access model accuracy.

1. **SIMPLE FORECASTING MODELS**

Initial forecasting began with four simple forecasting models; mean, naïve, seasonal naïve and drift. As can be seen below from figure 3.1, none of these

models were particularly effective in forecasting retail alcohol sales. I used the RMSE (the square root of the mean squared error) and MAPE (mean absolute percentage error) to compare the models (figure 3.2) . For the training dataset, the seasonal naïve model performed best. The naïve model, however, performed best on the test data and accordingly I would rank as the best of the simple forecasting models.



A picture containing diagram

Description automatically generated

Fig 3.1 – Simple Forecasting Models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Training** | | | **Testing** | | |
| **Model** | **RMSE** | **MAPE** | **Rank** | **RMSE** | **MAPE** | **Rank** |
| Mean | 978.8 | 29.1 | 4 | 2827.6 | 47.4 | 4 |
| Naive | 462.6 | 13.2 | 2 | 484.3 | 6.7 | 1 |
| Seasonal Naive | 129.7 | 3.7 | 1 | 1045.3 | 17.5 | 3 |
| Drift | 463.7 | 12.9 | 3 | 670.9 | 9.8 | 2 |

Fig 3.2 – Simple forecasting model comparison

1. **REGRESSION MODELS**

I used multivariate regression models to forecast the retail alcohol sales. The initial model, model 0, included all the variables; trend, season, GDP and CPI. However, I found that trend was not statistically significant with a p value of 0.624 and so I removed it for model 1.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Trend** | **Season** | **GDP** | **CPI** | **F** |  |
| Model 0 | Y | Y\*\*\* | Y\*\*\* | Y\*\*\* | 4559 | 0.996 |
| Model 1 |  | Y\*\*\* | Y\*\*\* | Y\*\*\* | 5510 | 0.996 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Trend** | **Season** | **GDP** | **CPI** | **F** | **R squared** |
| Model 0 | Y | Y\*\*\* | Y\*\*\* | Y\*\*\* | 4559 | 0.996 |
| Model 1 |  | Y\*\*\* | Y\*\*\* | Y\*\*\* | 5510 | 0.996 |

Fig 4.1 – Regression models parameter details

MODEL 1

The R squared remained very high with a value of 0.996. This indicates that the fluctuation in retail sales can be explained by the independent variables of season, GDP and CPI. Model 1 performs much better on the training data than on the testing data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Training** | | **Testing** | |
|  | **RMSE** | **MAPE** | **RMSE** | **MAPE** |
| Model 1 | 55.9 | 1.6 | 501.4 | 7.5 |

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Fig 4.2 – Regression model 1 Fig 4.3 – Regression model 1

1. **ARIMA MODELS**

After examination of the data using RStudio’s unitroot functions, I established that seasonal differencing was required.

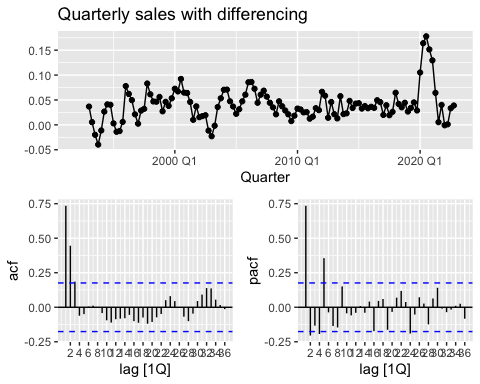


Fig 5.1 – Quarterly sales with differencing

I then used a number of ARIMA models for univariate forecasting. Using RStudio’s auto ARIMA, I determined that the best model was ARIMA (0,1,1)(0,1,2). Figure 5.3 displays the three ARIMA models while figure 5.4 includes only the best performing ARIMA (0,1,1)(0,1,2).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Training** | | **Testing** | |
| **ARIMA Model** | **RMSE** | **MAPE** | **RMSE** | **MAPE** |
| arima011210 | 50.6 | 1.4 | 696.8 | 11.6 |
| arima012110 | 53.9 | 1.5 | 718.3 | 12.1 |
| Arima011012 | 48.5 | 1.4 | 686.4 | 11.5 |

Fig 5.2 – Comparison of ARIMA models

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Fig 5.3 – Forecasting with ARIMA models Fig 5.4 - auto ARIMA (011,012)

1. **EXPONENTIAL SMOOTHING MODELS**

I used four exponential smoothing models; Simple Exponential Smoothing (SES), Holt-Winters, Holt-Winters Damped and Exponential smoothing model with trend and seasonality (auto). The ETS (auto) preformed the best when viewing the training dataset but the Holt model performed better for the testing data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Training** | | **Testing** | |
| **Model** | **RMSE** | **MAPE** | **RMSE** | **MAPE** |
| SES (ANN) | 382.2 | 9.4 | 763.3 | 9.6 |
| Holt (AAN) | 339.7 | 7.7 | 598.9 | 11.2 |
| Damped (AAdN) | 339.3 | 7.9 | 1138.5 | 20.7 |
| ETS (auto) | 69.8 | 1.4 | 763.3 | 13.1 |

Fig 6.1 – Model comparison

Chart

Description automatically generated

Chart, histogram

Description automatically generated

Fig 6.2 – Exponential smoothing models Fig 6.3 – ETS model

1. **MODEL SELECTION**

I then proceeded to compare the best preforming model from each category. I selected the naïve model, regression model 1, ARIMA (0,1,1)(0,1,2) and Holt AAN. While the ARIMA model performed best on the historic training data, the regression model performed better on the testing data. Accordingly, I believe that the regression model, using variables of season, GDP and CPI, is the best model for forecasting alcohol sales.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Training** | |  | **Testing** | |  |
| **Model** | **RMSE** | **MAPE** | **Rank** | **RMSE** | **MAPE** | **Rank** |
| Naïve | 432.3 | 13.3 | 4 | 629.1 | 10.9 | 4 |
| Regression | 53.6 | 1.6 | 2 | 376.6 | 5.2 | 1 |
| ARIMA | 43.3 | 1.3 | 1 | 459.6 | 5.9 | 2 |
| Holt | 270.8 | 7.5 | 3 | 546.2 | 8.8 | 3 |

Fig 7.1 – Comparison of best performing models

Chart, bar chart, histogram

Description automatically generated

Figure 7.2 – Comparison of best performing models

1. **CONCLUSION**

To conclude, the regression model did the best in forecasting retail alcohol sales. This is significant as it confirms that GDP and CPI play a very large factor in the fluctuation in alcohol sales. It is clear however that there are many other very relevant factors, for example the outbreak of the covid-19 pandemic. As can be seen from figures X below, when the shutdown of bars and restaurants occurred, naturally retail sales of alcohol increased to compensate, however, the pandemic caused a contraction in GDP. Generally, as seen from the regression the correlation between GDP and alcohol sales was positive, however, during the pandemic this became an inverse relationship.

Chart, line chart

Description automatically generatedChart

Description automatically generated

Fig 8.1 – Covid-19 sales increase Fig 8.2 – Covid-19 GDP decrease

**REFERENCES**

Rob J. Hyndman and George Athanasopoulos. Forecasting Principles and Practice, 3e, Otexts, Australia 2021.

Arora, Tanvi; Chandna, Rajat; Conant, Stacy; Sadler, Bivin; and Slater, Robert (2020) "Demand Forecasting In Wholesale Alcohol Distribution: An Ensemble Approach," SMU Data Science Review: Vol. 3: No. 1, Article 7.

<https://scholar.smu.edu/datasciencereview/vol3/iss1/7>

US Retail Sales: Beer, Wine and Liquor Stores [https://fred.stlouisfed.org/series/MRTSSM4453USN#0](https://fred.stlouisfed.org/series/MRTSSM4453USN)

US Gross Domestic Product

<https://fred.stlouisfed.org/series/NA000334Q>

Consumer Price Index for All Urban Consumers: Alcoholic Beverages in U.S City Average.

<https://fred.stlouisfed.org/series/CUUR0000SAF116#0>