

**FR IDM Process Overview**



|  |  |
| --- | --- |
| Project |  |
| Company | eClerx Services Ltd. |
| Prepared by | Dipika Patil & Saravana Ayyappa |

Contents

[1 Process 2](#_Toc19550970)

[**1.1** **Exploratory Data Analysis and data cleaning** 2](#_Toc19550971)

[**1.2** **Transformation of independent variables** 2](#_Toc19550972)

[**1.3** **Transformation of the dependent variable** 3](#_Toc19550973)

[**1.4** **Final Forecast** 4](#_Toc19550974)

[**1.5** **Verification of Model** 4](#_Toc19550975)

[**1.6** **Scenario Rerun**- 5](#_Toc19550976)

[2 Seasonality/business flags 5](#_Toc19550977)

[3 Consumer Visits: 6](#_Toc19550978)

[4 Consumer Online Units: 8](#_Toc19550979)

[5 Consumer Calls: 9](#_Toc19550980)

[6 SB Online Units: 11](#_Toc19550981)

[7 SB Visits: 13](#_Toc19550982)

[8 SB Calls 15](#_Toc19550983)

[9 SB Offline Units 17](#_Toc19550984)

1. Process
   1. **Exploratory Data Analysis and data cleaning**

**Univariate of dependent as well as independent variables is done to test:**

* Quantity and quality of data (check for missing values)
* Mean & Variance
* Outliers
* Inherent trends and seasonality

*Bivariate between dependent and independent as well as independent and independent variables to***:** Test for correlation between independent variables

**Treatment of missing values:**

* First check the reason for missing values, depending on which choose the treatment
* If there are sufficient data points, missing values can be removed
* If the variance is low, using mean is a good idea
* Mean cannot be used in the case of high variance

**Treatment of outliers:**

* First find if it needs to be treated. If they are intended to be high, then it should be left as it is. But knowing the outliers will help later in flagging them.
* (Mean ± 2Variance) is a commonly used way to find outliers. Other ways can be percentile (1, 5 depending on the need)
  1. **Transformation of independent variables**

**Lag**

If the whole effect of the independent variable is seen after a lag, the independent variable can be shifted in accordance to the trend. Eg. The effect of FSI (typically sent out on Thursdays), can be felt only after a couple of days. In this case lag transformation can be used.

**Allocation**

If the effect of an independent variable varies over a period of time i.e the effect is not same for the next couple of weeks, allocation can be used to get better correlation. Eg. The effect of a TV Ad will vary across the next few weeks (more in the initial first 2 weeks). We can allocate 30%, 30%, 20%, and 20% to weeks 1-4 in such case.

**Adstock:**

* This type of transformation is used when a long term effect of the independent variable is expected.
* For instance, TV ads and brand image has a long term effect even after the ad has been taken off
* The effect is diminishing: for instance, if adstock is 50%, the effect on Week1 will be 100%, week2 – 50%, week3 – 25%, week4 - 12.5% and so on
* Generally adstock is not favored unless there is a business sense, as there can be an interaction between vehicles and there could be chances of double counting
* Adstock must never be applied f or action vehicles

**Smoothing:**

* It is done to reduce large variations in data in cases where the variation in the independent variable is large, but the dependent variables’ variation is less.
* Eg, when the effect of a particular variable was high and seemed to reach a saturation and hence it was made low the following week (with not much effect on the dependent variable)
* Smoothing can be done in such cases by taking log functions, box-cox transformation etc.
  1. **Transformation of the dependent variable**

**Augmented Dickey-Fuller Test**

* While using ARIMA/ARIMAX to build a forecast model, one needs to make sure that the time series is stationary
* This test is used to check whether the time series is stationary.
* Here the null hypothesis (Ho) is that the data is non-stationary and the alternative hypothesis (H1) is that the data is stationary.
* So, if the time series is non-stationary we can make it stationary by calculating ‘d’ using inverse ACF

**Partial Auto Correlation Function**

* This test is done to calculate the p
* In Auto correlation function we take the correlation of current week value with previous week values by keeping the effect of independent variables whereas in PACF we remove the effect of independent variable to get the accurate picture
* So, we find correlation and if that correlation value is greater than the threshold value then we take the effect of that week considerable and then p is 1 and we take previous week values as another variable in the model

**Auto Correlation Function**

* If the correlation value has trend in consecutive lags then the data is non-stationary
* On the other hand if there is exponential decrease in the correlation from lag 0 and soon it dies down then the data is called as stationary

**Inverse ACF**

* This is used to get the d i.e. the degree of differencing that is needed to make the data stationary
* Same process as in ACF that we see the correlation value and till the lag it is significant i.e. above threshold we take that as d

**Estimation**

* Here to check whether the error is correlated or not we are finding the Auto correlation for residuals and if the p value comes less than 5% then it means the residuals are correlated otherwise they are not
* By ACF we can find the q if the residuals(error) are correlated with previous weeks
  1. **Final Forecast**
* After creating the final dataset by accommodating for the seasonality and business changes through flags, the dataset is fed into R
* For most models, ARIMA/ARIMAX has been used.
* Now we run various iterations by trying different combinations of the variables depending upon their p values
* The combination where all the independent variables are significant and the MAPE is minimum is used to forecast the dependent variable
  1. **Verification of Model**

1. **MAPE(**Mean Absolute Percentage Error**)**

* Here we find the error for previous data by taking the forecasted values
* Then take the mean of the absolute value of the error
* If the mean error value is less than 5% then the model is good otherwise its bad, rechecking is required

1. **In sample – Out sample MAPE**

* Remove the last 5 data points i.e. the actual values and without them the forecasting is done – This forecasted value for extra 5 data points is the out-sample
* For first 195 data points it is called the in-sample*(Assuming 200 data points in the total)*
* MAPE is calculated for both and it has to be less than 5% for model to get accepted

***Apart from these two there are three different tests available***

* **R squared –** The ratio of explained variation / total variation. R squared is a measure of how well the model is able to predict the variation in the dependent variable.
* **AICC**
* **SBC**

While R square is used for model verification, AICC and SBC is used for comparing two models with different results

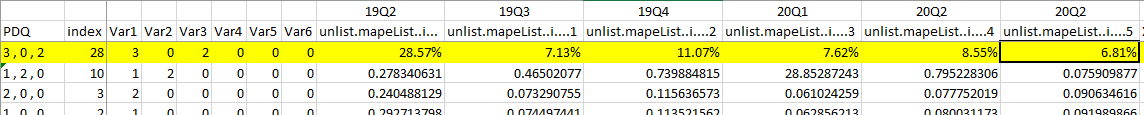
* 1. **Scenario Rerun**-
* There are two major inputs MARCOM spend and Pricing
* MARCOM spends are in the direct relation to the Y that is the dependent variable whereas the Pricing has the inverse relation that is with the increase in pricing there is a decrease in Y
* So in scenario rerun different combinations of spends and pricing are tried to get to different values of Y the dependent variable
* And out of these the best one is selected

1. Seasonality/business flags
2. **Labor day**- is a public holiday in **France** that celebrates workers' contributions to society. The holiday is celebrated annually on May 1.
3. **Black Friday and Cybermonday** - marks the start of the Holiday sales in the US. In France, it started in 2015 with only big companies sending out “Black Friday” or “Cyber” sales over the internet or even coupon through the mail. It was still discreet though.
4. **Christmas**- seasonal flag, an annual festival commemorating the birth of Jesus Christ, observed primarily on December 25. Government offices, organizations, schools and many businesses are closed on account of Christmas holiday.
5. **New** **Year**- seasonal flag, beginning of a new calendar year. Government offices, organizations, schools and many businesses are closed on account of New Year holiday.
6. **July Black Friday**- seasonal flag, initiative taken by Retail industry as researchers has mentioned that July is typically a down time in retail. Amazon has started the July Black Friday in 2015 by hosting it’s first-ever “Prime Day” shopping event. To include this flag, we check the trend and pattern of data at end of Q2.
7. **Good Friday** – Christian holiday commemorating the crucifixion of Jesus and his death at Calvary
8. **Easter** –a festival and holiday commemorating the resurrection of Jesus from the dead
9. **Victory Europe**- national holiday in France, it is the day that commemorates the end of World War II.
10. **Ascension Day -**marks the day that Jesus ascended to heaven following his crucifixion and resurrection, according to Christian belief. It is the 40th day of Easter and is ten days before Pentecost Sunday. It is a public holiday in France.
11. **Whit Monday-**Pentecost Monday, or Whit Monday, is the day after Pentecost. Many Christians believe that on Pentecost, the Holy Spirit descended on Jesus Christ's disciples. Pentecost Monday was a public holiday in France
12. **BTS –** back to school.
13. Consumer Visits:

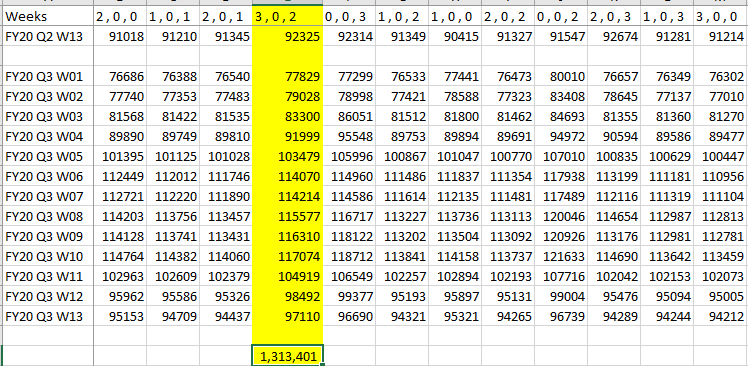
Variable\_List

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Variables | Variable \_Description | Notes | Coefficient |
|  | Fiscal\_Week | Mention of Fiscal Year, Quarter and Week |  |  |
| Visits | Actual\_Visits |  |
| Price | PRU | Price/HedgeRate |  | -0.0098282 |
| Marcom | LF | Combination of AFF+PL+OAL+SearchL |  | 0.0441072 |
| UF | Combination of OAU+SearchU+SMU |  | 0.1593755 |
| Flags | Pre\_BF | Marked for 18Q4 W3 and 19Q4 W3 | An explicit spike seen for 18Q4 and 19Q4 | 0.0592181 |
| BF | Marked for 18Q4 W4 and 19Q4 W4 | 0.0244169 |
| Christmas\_NY | Marked for 17Q4 W9-W10, 18Q4 W8-W9 and 19Q4 W9-W9 | Both the weeks performing almost similar as well on lower side | -0.0304241 |
| Good\_Friday | Marked for 17Q1 W8, 18Q1 W10, 19Q1 W8, 20Q1 W11 |  | 0.0095243 |
| Labour\_Day | Marked for 17Q2 W1 , Q1 W13 for FY18,19 and 20 |  | 0.0026816 |
| Victory\_Europe | Marked for 17Q2 W2 , Q2 W1 for FY18,19 and 20 | A slight dip observed in visits for 18 and 19 | -0.0001943 |
| Ascension\_Day | Marked for 17Q2 W1 , 18Q2 W3, 19Q2 W1 and 20Q2 W4 | Dip observed in visits | -0.0271125 |
| Whit\_Monday | Marked for 17Q2 W3, 18Q2 W5, 19Q2 W3 and 20Q2 W6 | Spike observed in visits | 0.0041655 |
| BTS\_1\_2 | Marked for Q3 W1 and W2 | Dip observed in visits , specially Q3 W2 | -0.0125011 |
| Q2\_Visits\_W5 | Marked for 19Q2 W5-9 and 20Q2 W6-W10 | Visits in these weeks seems to be on higher side than rest of the non-seasonal weeks | 0.0176644 |
| Q2\_Pre\_JBF | Marked for 19Q2 W10 and 20Q2 W10 |  | 0.0341441 |
| Q2\_JBF | Marked for 19Q2 W11 and 20Q2 W11 |  | 0.0178062 |

Index 28(3,0,2) has lowest Validation Mape(20Q2) = 6.81%



**Forecasts**:



**R code:**

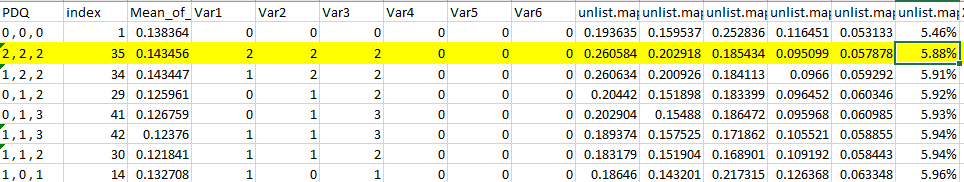


1. Consumer Online Units:

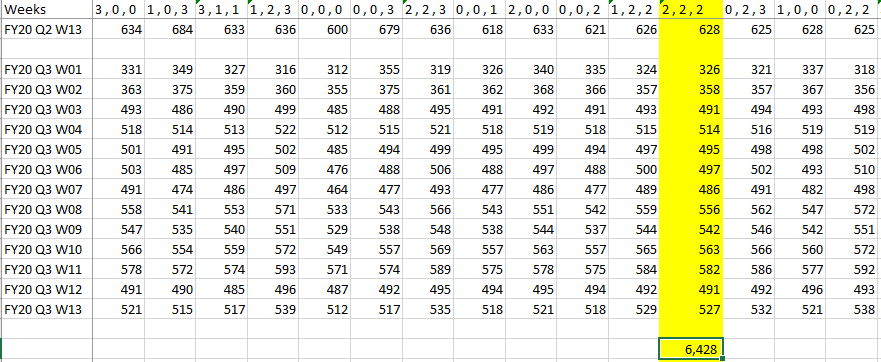
Variable List

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Variables | Variable \_Description | Notes | Coefficient |
|  | Fiscal\_Week | Mention of Fiscal Year, Quarter and Week |  |  |
| Online\_Units | Actual\_Visits |  |  |
| Price | PRU | Price |  | -0.11598119 |
| Marcom | LF | Combination of AFF+PL+OAL+SearchL |  | 0.01444911 |
| UF | Combination of OAU+SearchU+SMU |  | 0.00691196 |
| Flags | Pre\_BF | Marked for 18Q4 W3 and 19Q4 W3 |  | 0.11501998 |
| BF | Marked for 18Q4 W4 and 19Q4 W4 |  | 0.11476904 |
| Christmas\_NY | Marked for 17Q4 W9-W10, 18Q4 W8-W9 and 19Q4 W9-W9 |  | -0.04360401 |
| Pre\_Good\_Friday | Marked for 17Q1 W7, 18Q1 W9, 19Q1 W7 and 20Q1 W10 |  | -0.02650838 |
| Post\_Good\_Friday1 | Marked for 17Q1 W8-9, 18Q1 W10-11, 19Q1 W8-9 and 20Q1 W11-12 | Marked for Good Friday weeks and week after because they had similar performance | -0.00974186 |
| Q1\_W6 | A consistent dip observed for Q1 W6 |  | -0.01517429 |
| Q1\_W4 | A consistent dip observed for Q1 W4 |  | -0.01655953 |
| Q4\_Spike | Marked for 17Q4 W13, 18Q4 W12 and 19Q4 W11 | A consistent spike observed for those weeks | 0.04477185 |
| Q1\_W2\_Peak | A consistent spike observed for Q1 W2 |  | 0.02226665 |
| BTS\_1\_2 | Marked for Q3 W1 and W2 | Dip observed in units , specially Q3 W2 | -0.0571005 |
| Q2\_W13\_Q3\_W1 | A consistent dip observed for Q2W13 and Q3W1 |  | -0.02275722 |
| Q1\_Q2 | Marked for Q2 W1-W2-W3 | WoW a pattern of Dip-Spike-Dip observed for Q2 W1-W2-W3 except 18Q2 W2 and 19Q2 W3 | -0.05150941 |
| Q2\_W5\_W6\_W7 | Marked for either Q2 W5, W6 or W7 depending on runrate of Q2 | Units for these weeks seem to be lower side | -0.07181637 |
| Q2\_high\_1 | Marked last weeks of Q2 (could be JBF weeks) | Spike observed in units for 19Q2 W9-W11 and 20Q2 W9-W11 | 0.08364819 |

Index 35(2,2,2) has lowest Validation Mape(20Q2) = 5.88%



**Forecasts**



**RCode**



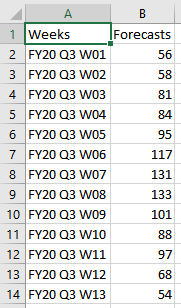
1. Consumer Calls:

Variable List

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Variables | Variable \_Description | Notes | Coefficient |
|  | Fiscal\_Week | Mention of Fiscal Year, Quarter and Week |  |  |
| Online\_Units | Actual\_Visits |  |  |
| Price | PRU Online | Price |  | -5.4815 |
|  | TMU Offline | Price |  | -1.2230 |
| Marcom | Affiliates | Affiliates Spends |  | 2.3715 |
| SM\_U | Social Media – Upper Funnel |  | 11.3688 |
| Flags | NY | Marked for 18Q4 W9 and 19Q4 W9 |  | -1.7613 |
| Annual Day | Marked for 19Q3 W5, W6 W7 and W8 |  | 7.5405 |
| BTS | Marked for 18Q3 W1, W2 & W3 and 19Q3 W1, W2 & W3 |  | 0.9432 |
| BF | Marked for 18Q4 W4 and 19Q4 W4 |  | 20.1123 |
| Good Friday | Marked for 18Q1 W10, 19Q1 W8 and 20Q1 W11 |  | -2.3180 |
| Post Easter | Marked for 18Q1 W12, 19Q1 W10 and 20Q1 W13 |  | -2.7041 |
| Assumption\_Day | Marked for 18Q3 W2 and 19Q3 W2 |  | -0.1892 |
| Q2\_W13 | Marked for 18Q2 W13, 19Q2 W13 and 20Q2 W13 |  | -6.3555 |
| BTS\_13\_3 | Marked for 18Q2 W13, 18Q3 W3, 19Q2 W13, 19Q3 W3 20Q2 W13, 20Q3 W1 and 20Q3 W2 |  | 5.2087 |
| Q4\_W13\_Q1\_W1 | Marked for 18Q1 W01, 18Q4 W13, 19Q1 W1, 19Q4 W13 and 20Q1 W1 |  | 8.1864 |
| W9\_W11 | Marked for 18Q1 W2, 19Q1 W9, 19Q1 W11, 19Q2 W9, 19Q2 W11, 19Q3 W9, 19Q3 W11, 19Q4 W10 19Q4 W12, 20Q1 W9, 20Q1 W11, 20Q2 W9, 20Q2 W11 |  | 2.6085 |
| Q1\_W6 | Marked for 18Q1 W06, 19Q1 W06 and 20Q1 W06 |  | -0.4194 |

Validation MSE (20Q2) = 12.21%

**Forecasts**



**RCode**

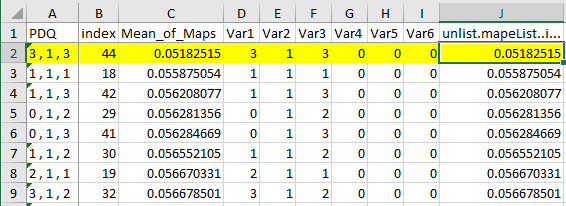


1. SB Online Units:

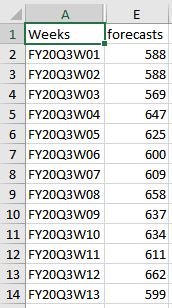
Variable List

|  |  |  |  |
| --- | --- | --- | --- |
|  | Variables | Variable \_Description | Coefficient |
|  | Fiscal\_Week | Mention of Fiscal Year, Quarter and Week |  |
| Online\_Units | Actual\_Visits |  |
| Price | PRU Online | Price | -0.127679211 |
| Marcom | Aff\_PL\_Search | Combination of AFF+PL+ Searche\_L & Search\_U | 0.011886905 |
| SM\_U | Social Media – Upper Funnel | 0.012458794 |
| Flags | Negative\_Spike\_OnlineU | The Negative Trend Observed during the weeks are Marked | -0.13428491 |
| Positive\_Spike\_OnlineU | The Positive Trend Observed during the weeks are Marked | 0.065902366 |

Index 44(3, 1, 3) has lowest Validation MAPE (20Q2) = 5.18%



**Forecasts**



**RCode**

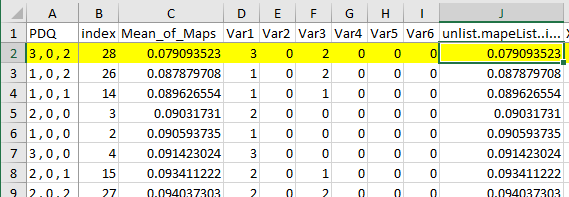


1. SB Visits:

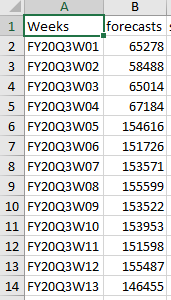
Variable\_List

|  |  |  |  |
| --- | --- | --- | --- |
|  | Variables | Variable \_Description | Coefficient |
|  | Fiscal\_Week | Mention of Fiscal Year, Quarter and Week |  |
| Visits | Actual\_Visits |
| Price | PRU Online | Price/Hedge Rate | -0.028831 |
| Marcom | OA\_L | Online Affiliates – Lower Funnel | 0.191407 |
| OA\_U | Online Affiliates – Upper Funnel | 0.104807 |
| Radio\_Newspaper | Combination of Radio & Newspaper | 0.016634 |
| Aff\_Search | Combination of Affiliates & Searche\_L | 0.005587 |
| SM\_U | Social Media – Upper Funnel | 0.040980 |
| Flags | Negative\_Spike\_Visits | The Negative Trend Observed during the weeks are Marked | -0.036684 |
| Easter | Marked for 19Q1 W9 and 20Q1 W12 | -0.003861 |
| Black\_Friday | Marked for 18Q4 W3, 19Q4 W4 and 20Q2 W10 & W11 | 0.0207530 |
| Christmas | Marked for 18Q4 W8 and 19Q4 W8 | -0.045472 |
| New Year | Marked for 18Q4 W9 and 19Q4 W9 | -0.044034 |
| Assumption\_Day | Marked for 18Q3 W1&W2, 19Q2 W13, 19Q3 W1 & W2 & 20Q3 W2 | -0.029553 |
| All\_Saints | Marked for 18Q3 W13 , 19Q3 W13, and 20Q3 W13 | -0.000220 |

Index 28(3, 0, 2) has lowest Validation MAPE (20Q2) = 7.90%



**Forecasts**:



**R code:**

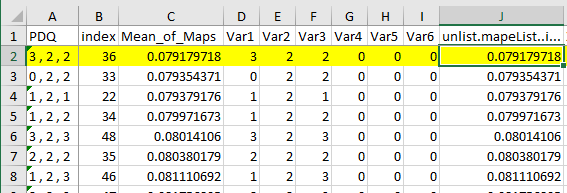


1. SB Calls

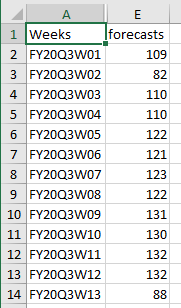
Variable List

|  |  |  |  |
| --- | --- | --- | --- |
|  | Variables | Variable \_Description | Coefficient |
|  | Fiscal\_Week | Mention of Fiscal Year, Quarter and Week |  |
| Online\_Units | Actual\_Visits |  |
| Price | PRU Offline | Price | -0.04595033 |
| Marcom | OA\_L | Online Affiliates – Lower Funnel | 0.070839182 |
| Newspaper\_Magazine | Newspaper & Magazines | 0.039206411 |
| Flags | Negative\_Spike\_Calls | The Negative Trend Observed during the weeks are Marked | -0.063970943 |
| Positive\_Spike\_Calls | The Positive Trend Observed during the weeks are Marked | 0.0359549145 |
| Assumption\_Day | Marked for 18Q3 W1&W2, 19Q2 W13, 19Q3 W1 & W2 & 20Q3 W2 | -0.082134918 |
| All\_Saints | Marked for 18Q3 W13 , 19Q3 W13, and 20Q3 W13 | -0.062459276 |
| Bastille\_Day | Marked for 19Q2 W11 & 20Q2 W11 | 0.013219156 |
| Christmas | Marked for 18Q4 W8 and 19Q4 W8 | -0.075673689 |
| New Year | Marked for 18Q4 W9 and 19Q4 W9 | -0.011016016 |
| Black Friday | Marked for 18Q4 W3, 19Q4 W4 and 20Q2 W10 & W11 | 0.0071642525 |

Index 36(3, 2, 2) has lowest Validation MAPE (20Q2) = 7.91%



**Forecasts**



**RCode**

****

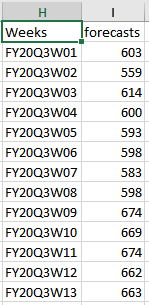
1. SB Offline Units

Variable List

|  |  |  |  |
| --- | --- | --- | --- |
|  | Variables | Variable \_Description | Coefficient |
|  | Fiscal\_Week | Mention of Fiscal Year, Quarter and Week |  |
| Online\_Units | Actual\_Visits |  |
| Price | PRU | Price | -73.87 |
| Marcom | Search\_L | Search – Lower Funnel | 97.31 |
| Search\_U | Search – Upper Funnel | 25.29 |
| Magazine | Magazine Spends | 32.09 |
| Flags | Easter | Marked for 19Q1 W9 and 20Q1 W12 | -27.09 |
| Labour Day | Marked for 18Q1 W13, 19Q1 W13 & 20Q1 W13 | -24.34 |
| Assumption\_Day | Marked for 18Q3 W1&W2, 19Q2 W13, 19Q3 W1 & W2 & 20Q3 W2 | -26.51 |
| Armistice Day | Marked for 18Q4 W2 & 19Q4 W2 | -11.86 |
| Christmas | Marked for 18Q4 W8 and 19Q4 W8 | -38.24 |
| New Year | Marked for 18Q4 W9 and 19Q4 W9 | -33.57 |

Validation MSE (20Q2) = 10.05%

**Forecasts**



**RCode**

