Marketing Mix Modelling (MMM) Methodology

Data Office – Data Science Team

December 2020

Contents

Background

A Marketing Mix Model (MMM) can from a high level be characterised as a statistical modelling technique that seeks to identify the relationship between your marketing spend in each individual marketing channel (eg paid online retargeting, social media, TV), and desired outcome or KPI, such as sales, website visits, and client acquisitions.

Methodology Overview

The proposed methodology for forecasting model selection has the following high-level steps:

- Model Base Form
- Model Attributes and Characteristics
- Data Requirements
- Model Evaluation
- Evaluation Metrics
- Advanced Features and Improvements

We next detail each of these components.

Methodology Detail

Model Base Form

In its base form, the model can be characterised as the following multiple linear regression:

$$S_t = \beta_0 + \sum_{i=1}^{n} \beta_i x_{ti}$$

where:

S_t – total sales at time t

β_0 – baseline sales, represented by brand loyalty or equity, in the absence of any marketing spend

x_{ti} – represents channel spend *i* (eg online paid, social, TV) by period (eg week, month)

β_i – coefficient or sensitivity of sales to channel x_{ti}

Model Attributes and Characteristics

The table below gives an overview of Nomad Foods Marketing Mix Model (MMM) characteristics. Please refer to ***Appendix – Time Series Forecasting Taxonomy*** for the detail of each of these terms.

Table – Attributes of Target Nomad Foods Marketing Mix Model (MMM)

Category	Nomad Foods MMM Attribute	Comment
Inputs and Outputs	Inputs - marketing spend by channel Outputs – revenue, conversion	
Endogenous vs Exogenous	Exogenous	The model will not depend on the revenue history itself
Regression vs Classification	Regression	The model will rely on various regression rather than classification techniques
Unstructured vs Structured	Structured	The data have systematic time-dependent patterns in a time series variable such as trend and/or seasonality
Univariate vs Multivariate	Multivariate	Multivariate model including as many channel metrics as possible
Single vs Multi Step	Multi-step	The business stakeholders may like to see forecasts over multiple horizons
Static vs Dynamic	Dynamic	The model is expected to be continuously re-trained or re-calibrated given changing nature of the business and revisions of analysts
Contiguous vs Dis-contiguous	Contiguous	Most of the data are expected to be relatively high quality and few missing values

Data Requirements

There following are the basic requirements for the data to build a model that is fit for purpose:

- **History** – sufficient history is needed to avoid overfitting
- **Granularity** – if there are requirements to predict metrics on divisional or product level, the input data should
- **Data quality** – data should have relatively few missing values or outliers

Model Evaluation Scheme

Below we present a typical regression time series forecasting model evaluation scheme.

Data Preparation and Pre-Processing

Some of the common data preparation or pre-processing methods include:

- Interpolation
- Outlier detection
- Hot encoding – converting categorical values to numerical values
- Differencing to remove a trend
- Seasonal differencing to remove seasonality
- Rescaling to normalise values

Split Data into Training and Test Sets

The data set should split the dataset into a train (calibration) and test set, for example 80% for training and 20% for testing, or out of sample prediction.

Feature Importance and Selection

In cases where were have many predictors or features, we may want to first filter out the features that explain the forecast best to to reduce overfitting and improve performance. Some of the common feature selection methods include:

- **Correlation and Mutual Information** – evaluating correlation and mutual information
- **SHAP score** – a framework to explain and visualise feature importance and impact
- **Recursive Feature Selection (RFE)** - RFE works by creating predictive models, weighting features, and pruning those with the smallest weights, then repeating the process until a desired number of features are left

Training and Calibration

The model candidate will fit on the training dataset. An important consideration on this step is to ensure that any coefficients used for data preparation are estimated from the training dataset only and then applied on the test set to avoid overfitting. This might include mean and standard deviation in the case of data standardization.

Forecasting

In this step, we preform out-of-sample prediction**,** or forecasting and compare to the test set directly or using walk-forward cross validation

Performance Evaluation

Finally, we calculate various performance metrics that compare the predictions to the expected values to evaluate goodness of fit of the model. Please see the next section **Evaluation Metrics** for further detail.

Model Evaluation and Goodness of Fit Metrics

For classical statistical methods, the following standard metrics can be used:

- **R-squared** - the R^2 (or R Squared) metric provides an indication of the goodness of fit of a set of predictions to the actual values. In statistical literature, this measure is called the coefficient of determination.
- **Mean Absolute Error (or MAE)** - Mean Absolute Error (or MAE) is the average of the absolute differences between predictions and actual values. It gives an idea of how wrong the predictions were.
- **Mean Squared Error (or MSE) -** much like the mean absolute error in that it provides a gross idea of the magnitude of error
- **Mean Squared Error (or MSE)** - Root Mean Squared Error (or RMSE) converts the units back to the original units of the output variable and can be meaningful for description and presentation, by taking the square root of the mean squared error
- **Back testing heuristics** – when back testing is applied, additional heuristics such as % time within accepted tolerance can be applied

The most appropriate metrics should be agreed upon with business stakeholders and end users.

Advanced Features and Extensions

Marketing Carry Over Effect or Lag

Not all marketing sees an immediate effect. Many if not most consumers today go through a decision-making or consideration phase that starts when the awareness is created until a decision is made to either purchase or not purchase.

The consideration phase is represented through a decay rate, which is the rate at which the marketing expenditure decays from one period to another.

The carry over effect can be represented as:

$$A_t = x_t + (\gamma * A_{t-1})$$

where:

A_t – total sales at time t

x_t – marketing channel spend at time t

γ – decay rate

An advanced model should account for this lag appropriately.

Appendix – Time Series Regression Model Types

Below is a typical model evaluation order structured in increasing complexity from classical to modern methods.

- **Baseline**
    - Single linear regression and Ordinary least squares (OLS)
    - Multiple linear regression
- **Autoregression**
    - ARMA for stationary data
    - ARIMA for data with a trend
    - SARIMA for data with seasonality
    - VAR/VARMA for multivariate time series
- **Exponential Smoothing**
    - Simple Smoothing
    - Holt Winters Smoothing
- **Linear Machine Learning**
    - Linear Regression
    - Ridge Regression
    - Lasso Regression
    - Elastic Net Regression
- **Nonlinear Machine Learning**
    - k-Nearest Neighbours
    - Classification and Regression Trees
    - Support Vector Regression
- **Ensemble Machine Learning**
    - Bagging
    - Boosting
    - Random Forest
    - Gradient Boosting
- **Deep Learning**
    - MLP
    - CNN
    - LSTM
    - Hybrids

Ideally, each model evaluation experiment should record results to a file so that multiple evaluation runs could be easily compared.

Note that experts often give deep learning methods lower value in this context as generally neural networks are poor at time series forecasting, but there is still a lot of room for improvement and experimentation in this area

References

- How to Develop a Skillful Machine Learning Time Series Forecasting Model - Jason Brownlee
- Feature Selection for Time Series Forecasting - Jason Brownlee
- 11 Classical Time Series Forecasting Methods - Jason Brownlee
- SHAP Feature Importance with Feature Engineering – Kaggle
- Dealing with Multicollinearity - Kaggle

Appendix - Time Series Forecasting Taxonomy

Inputs vs. Outputs

- **Inputs** - historical input data provided to the model in order to make a future forecast, also known as predictors, independent variables or features
- **Outputs** - prediction or forecast for a future time step beyond the data provided as input, also known as dependent variable, target variable or label

Endogenous vs. Exogenous

- **Endogenous** - input variables that are influenced by other variables in the system and on which the output variable depends. See also Multicollinearity in the Appendix
- **Exogenous** - input variables that are not influenced by other variables in the system and on which the output variable depends.

Regression vs. Classification

- **Regression**- forecast a numerical quantity
- **Classification** - classify as one of two or more labels

A regression problem can be reframed as classification and a classification problem can be reframed as regression. Some problems, like predicting an ordinal (categorical) value, can be framed as either classification and regression. It is possible that a reframing of your time series forecasting problem may simplify it

Unstructured vs. Structured

- **Unstructured** - no obvious systematic time-dependent pattern in a time series variable.
- **Structured** - systematic time-dependent patterns in a time series variable (e.g. trend and/or seasonality)

We can often simplify the modelling process by identifying and removing the obvious structures from the data, such as an increasing trend or repeating cycle. Some classical methods even allow you to specify parameters to handle these systematic structures directly

Univariate vs. Multivariate

- **Univariate** - one variable measured over time.
- **Multivariate** - multiple variables measured over time.

In terms of inputs / outputs, we can also consider the following breakdown:

- **Univariate and Multivariate Inputs** - one or multiple input variables measured over time.
- **Univariate and Multivariate Outputs** - one or multiple output variables to be predicted.

Single-step vs. Multi-step

- **One-Step** - forecast the next time step
- **Multi-Step** - forecast more than one future time steps

The more time steps to be projected into the future, the more challenging the problem given the compounding nature of the uncertainty on each forecast time step.

Static vs. Dynamic

- **Static** - a forecast model is fit once and used to make predictions.
- **Dynamic** - a forecast model is fit on newly available data prior to each prediction.

Contiguous vs. Dis-contiguous

- **Contiguous** - observations are made uniform over time
- **Dis-contiguous** - observations are not uniform over time

The lack of uniformity of the observations may be caused by missing or corrupt values, and in such cases specific data formatting may be required when fitting some models to make the observations uniform over time.

Appendix - Time Series and Regression Glossary

Decomposition Modelling

Decomposition modelling involves breaking a time series into the components of Trend, Seasonality, Cyclicity and Irregularity, described below.

Trend

Persistent over a relatively long period of time, the trend is the overall increase or decrease of the series during that time.

Seasonality

Seasonality is the presence of variations that occur at specific regular intervals; it is the component of the data and series that experiences regular and predictable changes over a fixed period.

Cyclicity

Cyclicity refers to the variation caused by circumstances, which repeat at irregular intervals. Seasonal behaviour is very strictly regular, meaning there is a precise amount of time between the peaks and troughs of the data; cyclical behaviour, on the other hand, can drift over time because the time between periods is not precise.

Irregularity

Irregularity is the unpredictable component of a time series — the 'randomness'. This component cannot be explained by any other component and includes variations which occur due to unpredictable factors that do not repeat in set patterns.

Autocorrelation

Sometimes also referred to as lagged correlation or serial correlation, autocorrelation refers to the degree of similarity between a given time series, and a lagged version of itself over successive time intervals. Mathematically it is defined as the standard Pearson correlation coefficient between S(t) and S(t-k) where S(t) is the value of the series at time t and is the lag.

Cross Correlation

Similar to autocorrelation, cross correlation represents the Pearson correlation coefficient between two time series S(i,t) and S(j,t-k), indexed by i and j where the latter is lagged by lag k.

Spurious Correlation

Spurious correlation is a mathematical relationship in which two or more events or variables are associated but not causally related. This can be due to either coincidence or the presence of a third, unseen factor, sometimes called a “common response variable”, “confounding factor”

Stationarity

A stationary time series is one in which several statistical properties — namely the mean, variance, and covariance — do not vary with time. This means that, although the values can change with time, the way the series itself changes with time does not change over time. Most classical time series forecasting methods require converting the time series into a stationary form to improve accuracy.

Mulitcollinearity

Multicollinearity (also collinearity) is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. In this situation, the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors.

Granger Causality Test

Granger Causality Test determines whether one time series will be useful in forecasting another. See more at Granger Causality

Dynamic Time Warping

Dynamic Time Warping (DTW) is a family of algorithms which compute the local stretch or compression to apply to the time axes of two timeseries in order to optimally map one (query) onto the other (reference). DTW outputs the remaining cumulative distance between the two and, if desired, the mapping itself (warping function). DTW is widely used e.g. for classification and clustering tasks in econometrics, chemometrics and general timeseries mining. For Python implementation see dtw-python package.