EDX - Capstone ML Parking Sales (Independent Module)

John Horbino

13/06/2020

Foreword

The data used in this report and the underlying R code is based of sales data from a real car parking facility in Australia. As such many of the fields have been de-identified and only a sample has been extracted for this report.

Introduction

To the general public, car parking is often an under looked piece of infrastructure that acts as the primary access point to many facilities. As technology has improved, many car parking businesses allow customers to book online in advance, benefiting both parties. The sales dynamics of online bookings are significantly different to the traditional drive-up experience as many other factors are introduced to the customers. Within this dynamic lies the complexity of long term trends and short term variations in car parking demand. The main goal of this report is to develop a machine learning algorithm that translates a lot of short term variation into business friendly factors. To achieve this the algorithm should be usable in a daily operational environment for performance attribution and also be scalable for future development.

A customer's choice of where to park and length of stay is driven by a commitment, or desire, to be somewhere. This can include commitments such as shopping, going on holiday or coming into work. Within these underlying commitments lies the different options available to a customer to get to their point of interest; not all modes of transport translate into the use of car parking. What often drives a customer to purchase a car parking product is the logistical convenience associated to their destination. What drives them away can be related to the price, especially against cheaper modes of travel such as ride share (Uber, Didi, Ola, etc) or public transportation (bus or train).

In general the customer demand for car parking products can be interpreted as a time series that can be broken down into:

1. Long term: A persistent source of commitment.
2. Short Term: Various time dependent factors.

To achieve the main goal of this report, machine learning methods will be used to assist in decomposing the demand time series. While the focus will be on the short term variations, a long term approximation will be develop so that it can be decoupled from the time series.

Data Sourcing, Cleanliness & Definitions

The data has been provided in the form of an R.Data workspace, initially sourced from the corporate data warehouse. As mentioned in the foreword, key business fields have been de-identified and will generally be referred to by ID, unless they are sufficiently general by name. As the source data has undergone various

ETL steps from the data warehouse, it is in a tidy format. Minimal transformations have been performed to make later calculations easier.

The sampled data is based on sales between 1 January 2017 and 30 May 2020. However the available data from Google Analytics only begins in late 2018. As such there is a calculation impact towards the training algorithm as the training must only be applied to the commonly available dates (will be discussed in "Methods - Key Steps").

Below are terms that will be used throughout the report:

- ReportLevel4Key - The ID relating to the car parking product.
- StayLengthName - Denotes the length of time a customers has chosen to occupy a car parking space.
- StayLengthKey - The ID relating to StayLengthName.
- ATV - Short for "Average Transaction Value" which relates to the average price of a product for a particular occupancy time.
- EDM - Short for "Electronic Direct Mail" is a form of digital marketing that uses email to target large groups of people.
- GA - Short for "Google Analytics" which is a source of website performance data, such as number of sessions or users that a website achieves.
- Centered Moving Average - Performs a moving average calculation with equal weightings to data points before and after the current point (often time series).

R Libraries

The following R libraries have been used in the analysis.

1. tidyverse - Forms the bread and butter of all data transformations.
2. caret - Primary use is to partition the data for training and test validation.
3. data.table - Used as an alternative to tidyverse for quicker and more efficient data transformations.
4. ggplot2 - General plotting.
5. zoo - Used alongside tidyverse and data.table to perform rolling/windowed calculations.
6. broom - Used alongside tidyverse to output tidy results from regression analysis.
7. lubridate - Used alongside tidyverse to parse out date components.

Dplyr, as part of tidyverse, can fully replace data.table, but this will have a material impact in execution time.

R Code Overview

The attached R code segments the code through outlines to make navigation easier. The progression of this report will follow closely to the R code.

Due to the number of explanatory variables, the custom function "applyTrainingResults" has been used to accept data in the same format as the training or test validation sets. It also requires an EDM related data frame as a second input in order to calculate the impact of marketing activities. This function outputs a list of different objects with the main result data frame "CoreResults". It also generates a suite of accuracy plots and tables to show a summary of the different product breakdowns.

Methods - Key Steps

Time series decomposition involves separating out the long term trend from the data in order to independently assess the seasonal and cyclical factors. As the focus of the report is to decompose the short term factors, an

approximation for the long term trend will be sufficient. In saying that, the long term trend plays a pivotal role in calculating the short term factors, which will be covered in later sections.

In the sample data provided, all the different combinations of ReportLevel4Key and StayLength-Key/StayLengthName act as different product groupings.

As mentioned in the "Data Sourcing, Cleanliness & Definitions" section the training must overcome the data availability difference between sales volume and Google Analytics. To do so, the training will prioritise on attributing factors that can use all of the sales volume data, then shrinking to accommodate for the marketing factors. This transition can be seen in the R code section "21 Calculate Non Seasonal Components". In general, time dependent seasonal factors will be trained on the full sales volume time series data. The data availability will be marked for each factor later in this report.

The general overview of the methods are as follows (performed for each product combination):

1. Approximate long term trend and remove it from the time series data.
    - Refer to R code section "10 Calculate long term trends".
2. Partition the data to training a test sets.
    - Refer to R code section "12 Partition primary data frame".
3. Calculate the impact of different seasonal and other non-seasonal short term factors.
    - Refer to the relevant sections in the R code section "20 Calculate Seasonality components".
4. Assess residuals and possibility of additional decomposition.
    - Refer to the R code section "29 Remaining Residuals".
5. Apply model to validation set for performance testing.
    - Refer to the R code section "40 Generate Training & Test Results".

Description of each factor explored in step 3 will be provided in the following sections.

Accuracy Measures

This report uses the following metrics to measure accuracy. Ideally these metrics should be as close to zero as possible. The goal of each factor is to decrease the remaining residual.

1. Root Mean Square Error (RMSE) - Calculates the average distance (scalar) of each sales volume prediction against what actually occured. Gives more weight to larger errors.
2. Mean Absolute Error (MAE) - Similar to RMSE but takes the absolute value of the residual instead of squaring and square rooting. Even weighting between large and small errors.
3. Mean Bias Error (MBE) - Similar to MAE but does not take the absolute value (vector quantity). If the resulting value is not 0 it suggested a model bias to under or over predict.

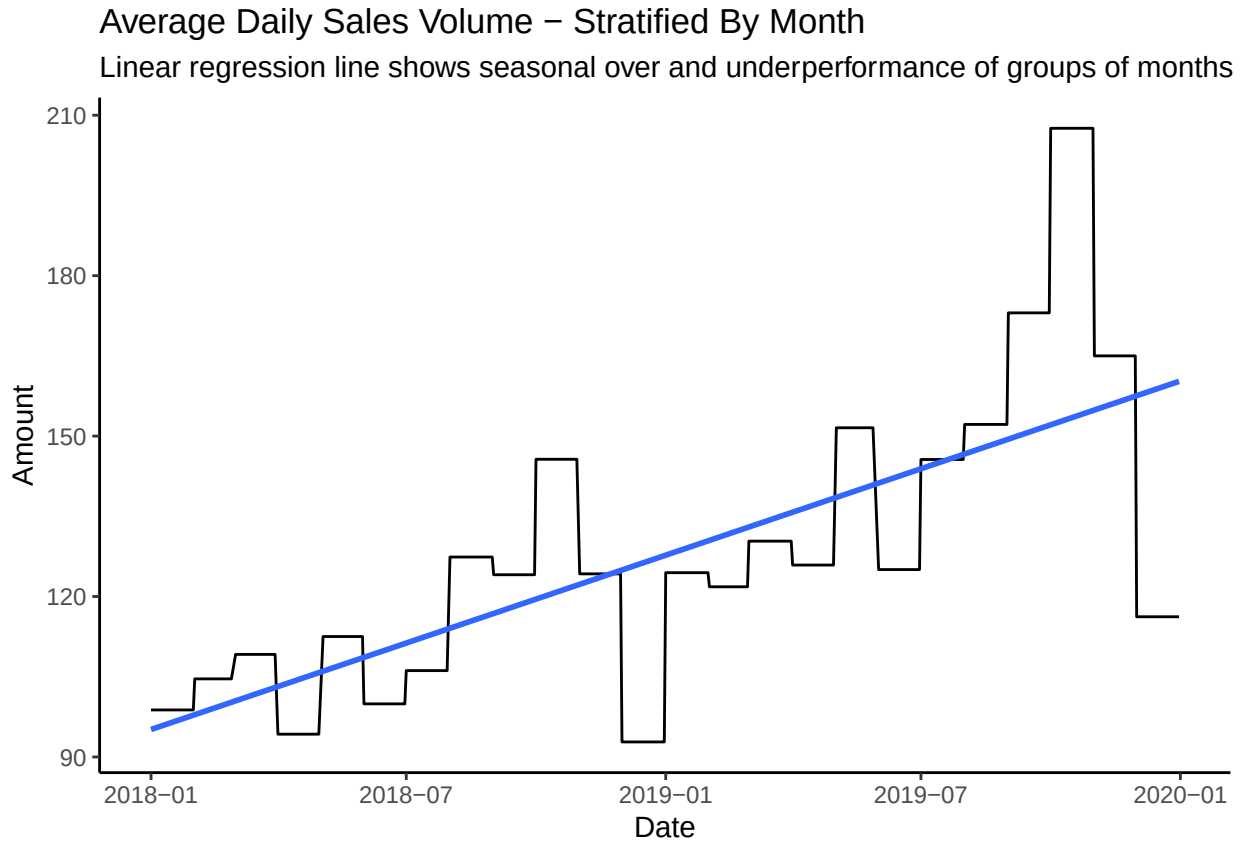
To try and contextualise the accuracy measures above, the following relative percentage versions have also been calculated by dividing their non percentage counterparts by the actual results. Resulting NAs, Inf and -Inf results are omitted. This can bring out potential biases in the resulting values which will be covered in the model accuracy results section.

4. RMSE Perc - specifically calculates the final RMSE of each group by the group's average.
5. MAE Perc.
6. MBE Perc.

Decomposition & Analysis

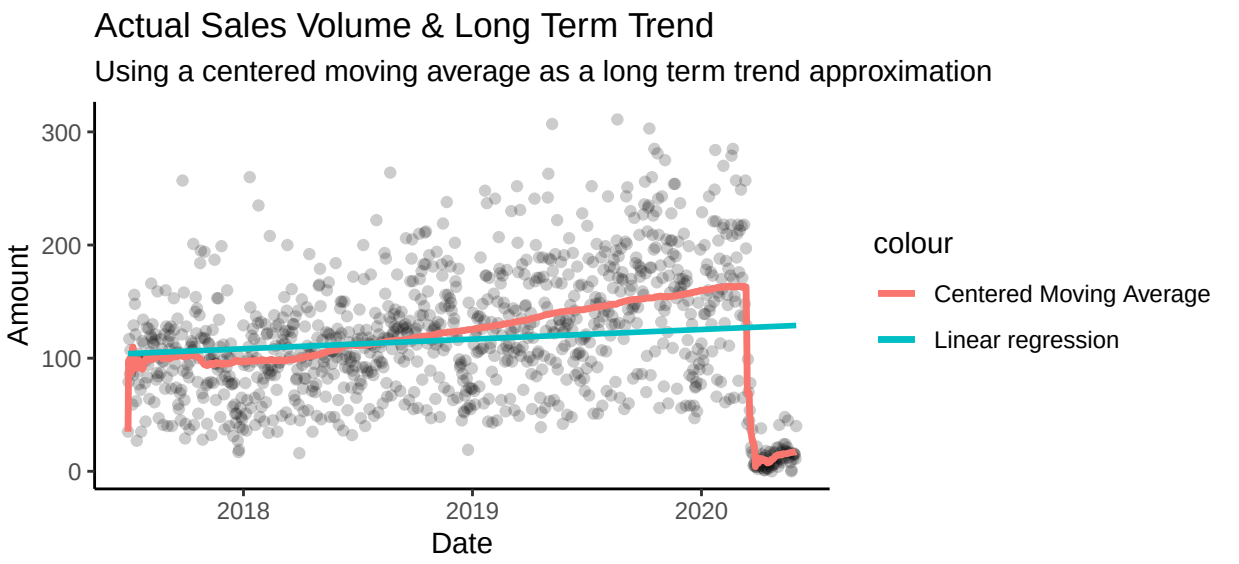
Long Term Trend

A common way to approximate a long term trend is to use a centered moving average that spans enough time points such that the variation from seasonal factors are sufficiently diluted. Sales related time series data would naturally have annual peaks and troughs as certain periods in the year perform better than others. An example would be heightened sales leading up to Christmas time. Applying this intuition to the sample data shows the relative over and under performance of some months when compared to a simple linear regression (to simulate a long term trend).



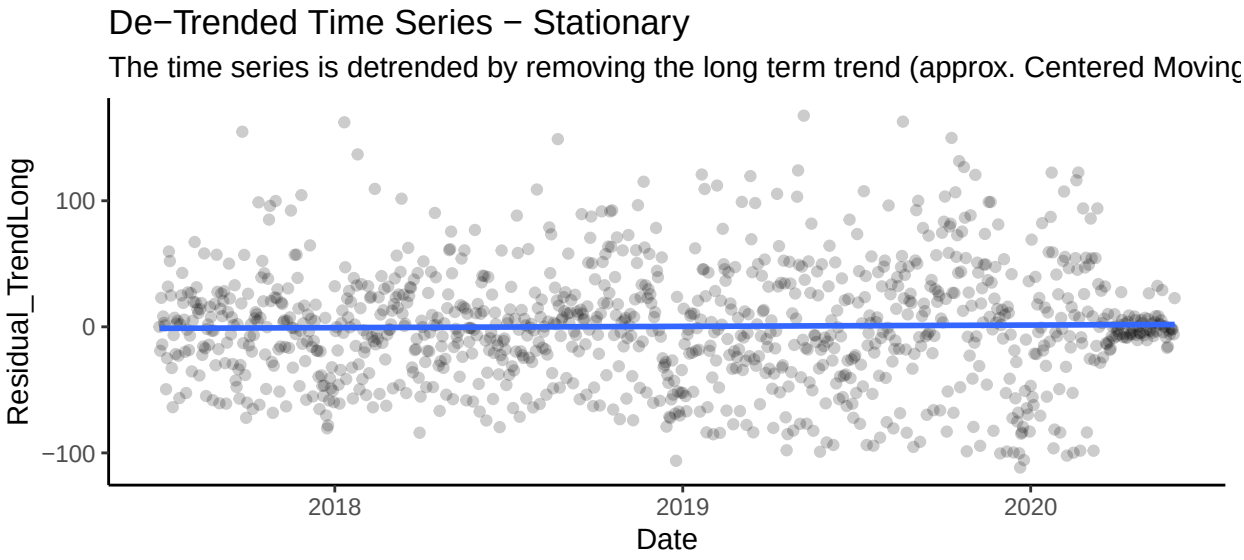
If annual seasonality is assumed to be the longest seasonal period in the time series, building a centered moving average that spans 364 days should produce a clear trend that is without any repeating local maxima or minima. The results of such a moving average can be seen below. It is quite clear that a 364 day span is sufficient to approximate a long term trend. The most notable dip in the data occurs during March 2020 when the COVID-19 virus made its initial impact in the industry.

To account for the clear shift in trends between pre-COVID and post-COVID, two windows of the centered moving average was calculated. The first spanning before 22-Mar-2020 while the second comes afterwards. This was necessary as the drop in sales due to COVID-19 is an irregular external shock to the industry with which could not be reasonably forecasted.



Short Term Trend - Partitioning

Once the long term trend is stripped out, the remaining residuals form a "stationary" time series data, which meaans that the data is centrally distributed around zero. Below shows this result.



The remaining data is then partitioned to prepare for short term seasonality and factor decomposition. The end goal of the exercise is to perform attribution on a daily basis and update the training model with as much data as possible in order to capture recent market shifts. To simulate this, an 80/20 split was chosen. Higher proportions can be used but can easily lead to over training, while lower proportions may leave too much data out to perform meaningful decomposition.

Short Term Trend - Overview

In every machine learning task, future use of training results must be considered. In the current context, scaling must be performed on the long term trend residuals so that the decomposed short term factors can

be applied to whatever magnitude of sales data. If decomposition is done on raw residuals, it will only be relevant in the sales volume magnitude performed during the training and will adversely over or under estimate results if the underlying levels change. The best example of this is the transition to post-COVID sales volume where the overall average level of sales is significantly lower than pre-COVID but the same seasonal patterns remain.

To scale the data, the residuals of each decomposition step are divided by the long term trend. By doing so, the impact of each decomposition step is always relevant to the underlying level of sales volume in the current context.

Each decomposition follows the general steps:

1. Generate a plot to visualise the residual percentage impact.
2. Calculate the grouped residual percentage impact and store in a separate data frame.
3. Apply the sensitivities to the long term trend to approximate the relevant value (real not percentage) of the decomposed factor.
4. Calculate new residuals after removing the impact as calculated in step 3.
5. Generate a plot as per step 1 but using the new residuals generated in step 4. The percentage impact should be very close to zero. If not then further assessment must be done with the calculation.

It is important to note that the magnitude of each factor's contribution is dependent on the order they are calculated but will always aggregate together to the same value. This is because the nature of the calculation is to average out the remaining residual after the prior calculation.

Heuristically, it can be thought as removing water from a well with differently shaped buckets - each removes a different amount of water but will always add to the same aggregate volume.

While the aggregate result remains the same, the differing impact of each factor would realistically affect decision making in a business environment. As such it is important to apply an order that resembles the business' understanding of how the factors relate to each other. In the context of the data, the decomposition generally starts from classically time dependent seasonality factors, to major marketing campaigns, and then to delayed effects of major marketing campaigns. More rigorous methods can be implemented in the future to average out the different impacts associated to different ordering.

Short Term Trend - Per Factor Decomposition

The different factors are listed below, in the order they are calculated. A short description and its link to business context is provided alongside a visualisation to demonstrate the predictive power.

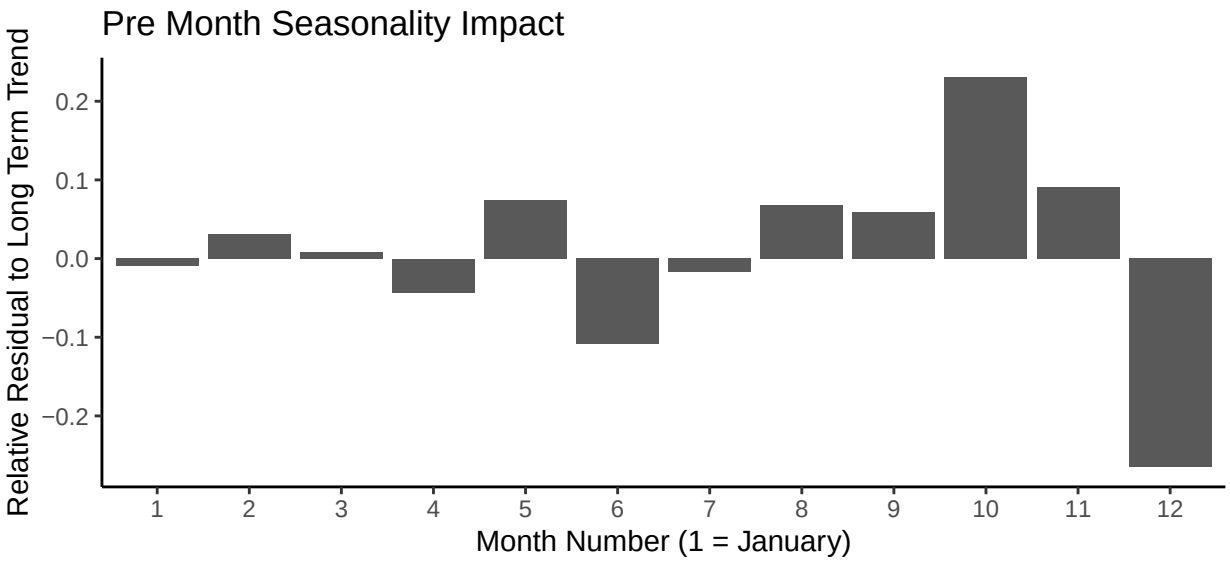
For additional calculation detail, please refer to the submitted R code.

1 - Seasonal Month

Uses all available training data.

This factor represents the residual variability of each month of the year. Calculated by taking the average percentage remaining residual per month.

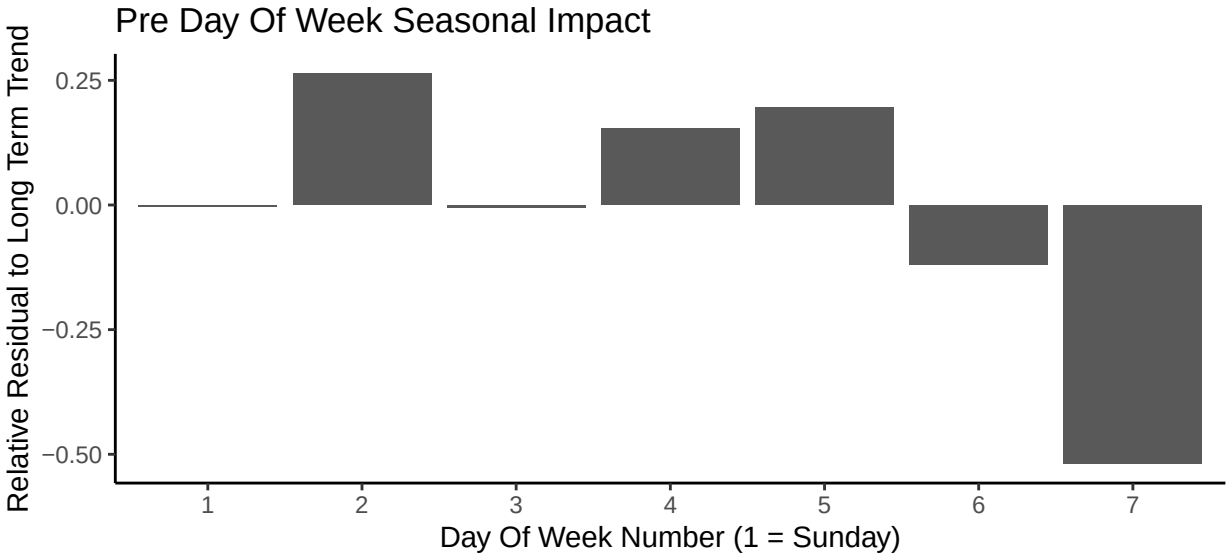
Some months experience an increased number of bookings while the opposite is true. The primary example are the months leading up to December as customers are booking for longer term holidays. The opposite is true for January and June which follow Christmas and Easter holidays respectively.



2 - Seasonal Day Of Week

Uses all available training data.

This factor represents the residual variability of each month of the year. Calculated by taking the average percentage remaining residual per day of week. This seasonality is heavily correlated with week day work schedules as there are many business related travel. As such Mondays to Thursdays experience much higher booking volumes whereas Friday and Saturday are the opposite.



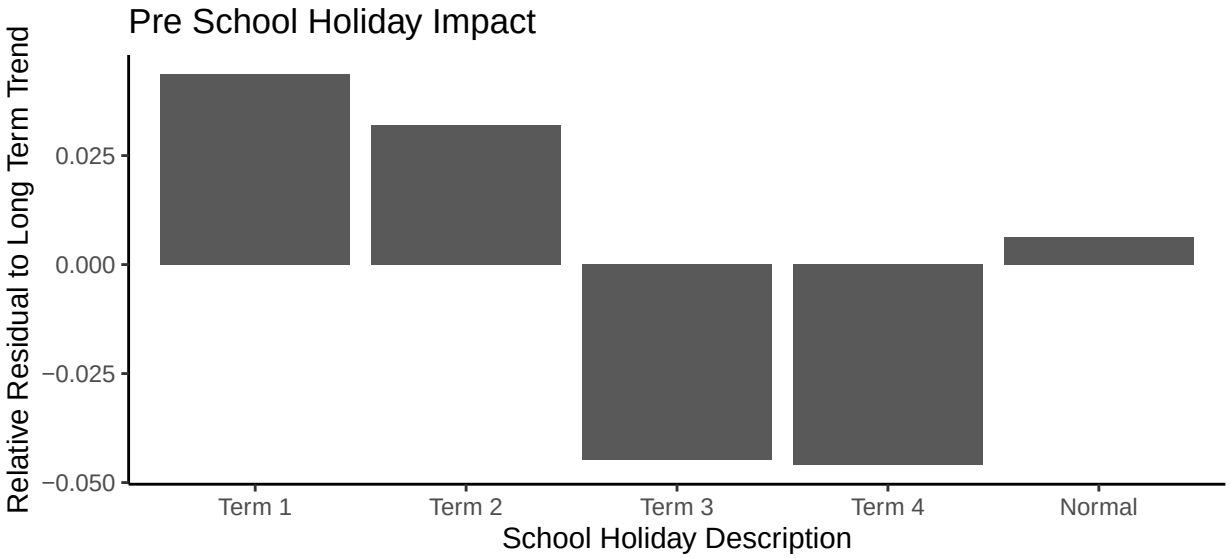
3 - Seasonal School Holiday

Uses all available training data.

This factor represents the residual variability of school holiday periods. The Australian school holiday periods are once per quarter on similar dates per year. Calculated by taking the average percentage remaining

residual per school holiday. This factor represents the impact of school students being at home and the higher propensity for families to go on holiday.

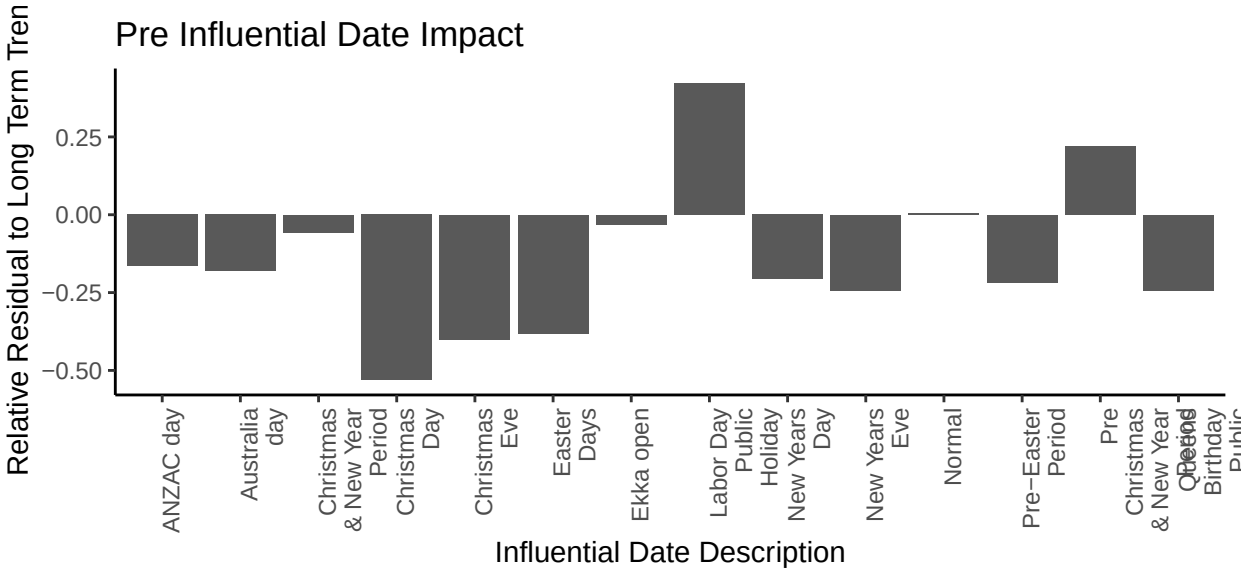
While the overall percentage residual per term is not as significant as the previous components, there is a clear difference between the first and last two terms.



4 - Seasonal Influential Dates

Uses all available training data.

This factor represents the residual variability of different major holiday periods. This dimension is a mixture of single day holidays, such as ANZAC and Australia Day, and major holiday periods, such as the Christmas period. Similar to school holidays, this factor also represents the purchasing behaviour of customers on these major holiday periods.

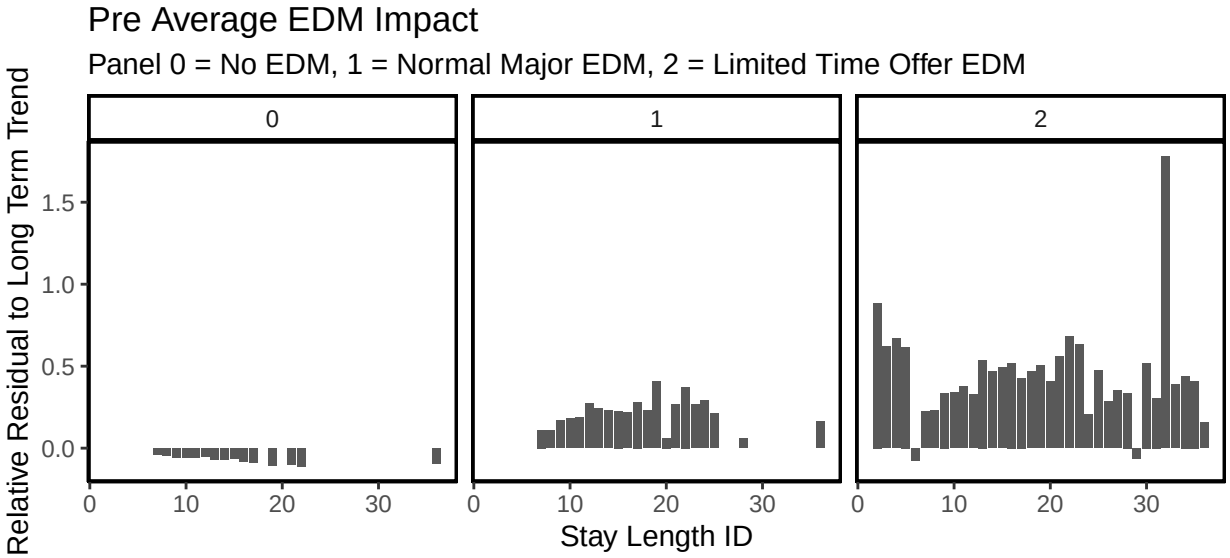


5 - EDM Major Average

Uses Google Analytics and related sales data (post October 2018).

This factor represents the average residual variability when EDMs are sent to customers.

When an EDM is sent, there is generally a consistent uplift in sales volume. This seems to be due to the perceived value for the customer as the parking EDMs usually come with discount offers.

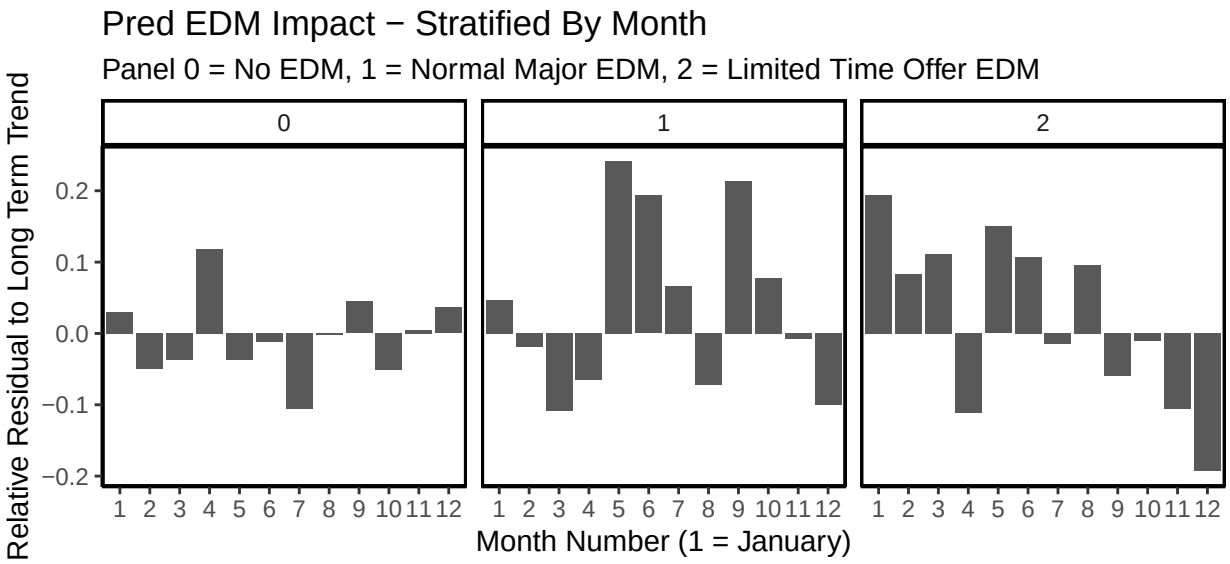


6 - EDM Major Month

Uses Google Analytics and related sales data (post October 2018).

This factor represents the monthly impact of the residual variability when EDMs are sent to customers.

The better performance of some months seems to line up with periods of high and low travel. For example, December performs the worst for EDMs, presumably because people have already booked well in advanced for the Christmas period. On the other hand the middle of the year performs the best, possibly due to heightened business travel and off-season holidays.

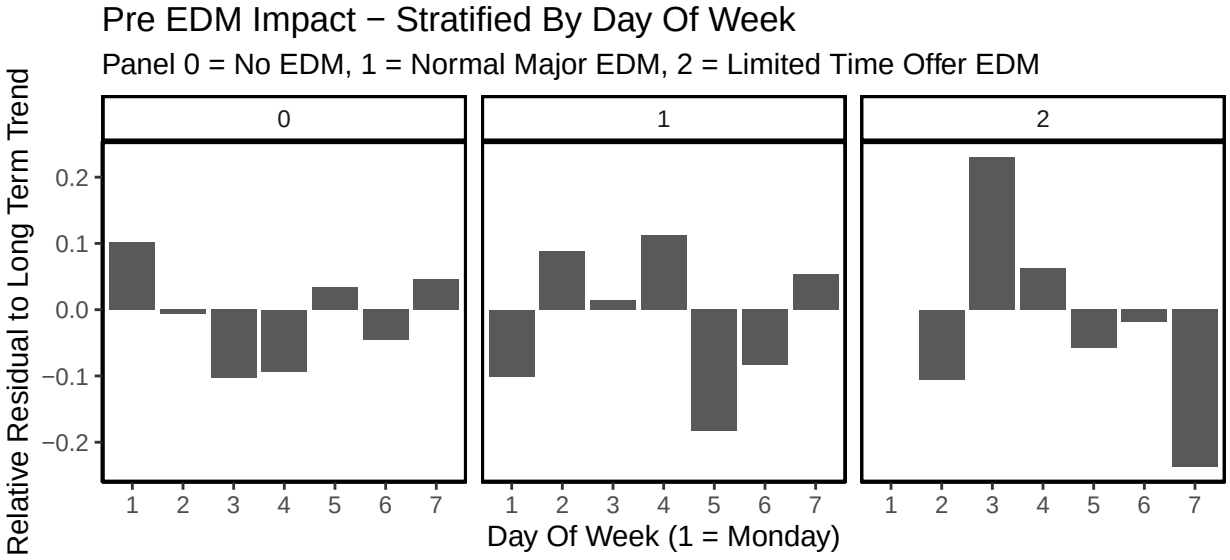


7 - EDM Major Day Of Week

Uses Google Analytics and related sales data (post October 2018).

This factor represents the day of week impact of the residual variability when EDMs are sent to customers.

Similar to the general day of week component, this component seems to benefit Monday to Thursday bookings, seemingly correlated with business travel behaviour.

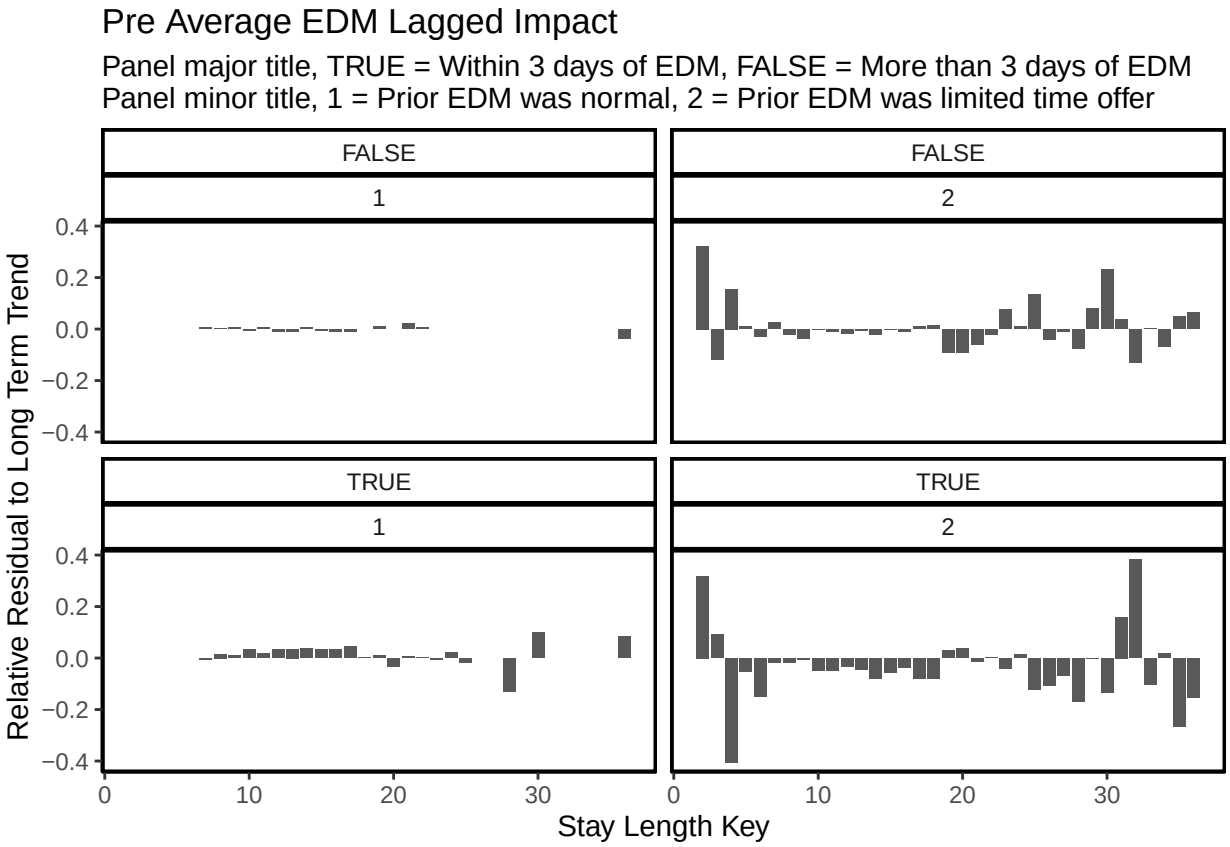


8 - EDM Major Lagged Average

Uses Google Analytics and related sales data (post October 2018).

This factor represents the delayed impact of the residual variability after EDMs are sent to customers.

While EDMs generally see significantly more bookings on the day they are sent out, a period of lower than trend performance is experienced for the following period. This is primarily due to the ability of EDMs to encourage people to book well ahead in time due to discount offers. The end impact results in shifting bookings ahead in time. What follows is usually a period of lower than trend sales volume as the market corrects itself. In general, EDMs result in a neutral sales volume position over a longer period.



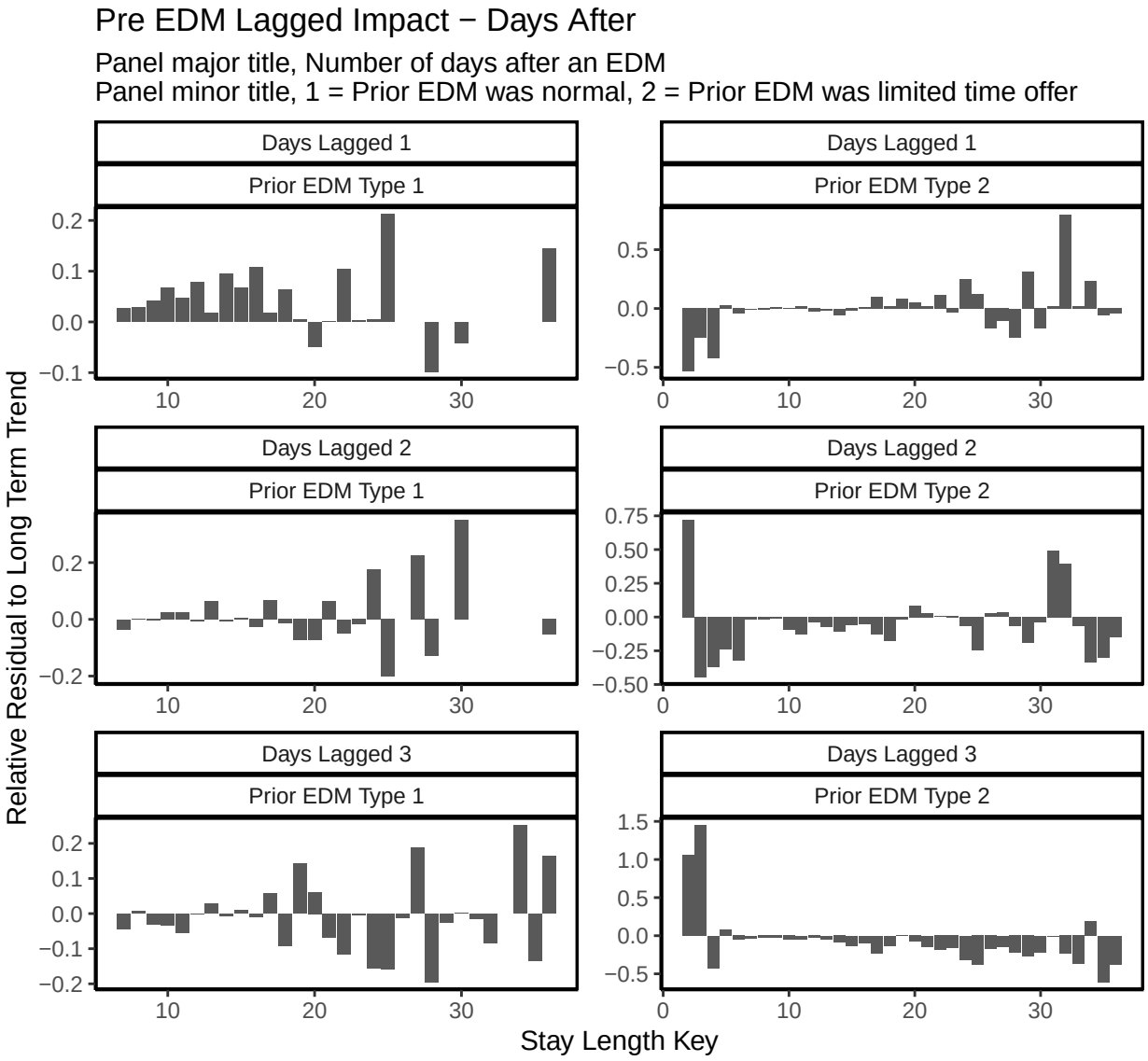
9 - EDM Major Lagged Days

Uses Google Analytics and related sales data (post October 2018).

This factor represents the impact of the number of days since an EDM is sent. The model currently factors in up to a three day lag. Additional days can be modelled in but the relative impact becomes less significant and relevant.

While the previous factor calculates an average level, this seeks to spread the delayed impact over the following days. For non limited time offer EDMs (those with a discount offer that lasts over 24 hours), the initial positive impact seems to carry forward for a few days after which it begins to show lower than trend performance. On the other hand, limited time offer EDMs seems to consistently see underperformance in the following days as the associated offer carries a time to purchase luxury.

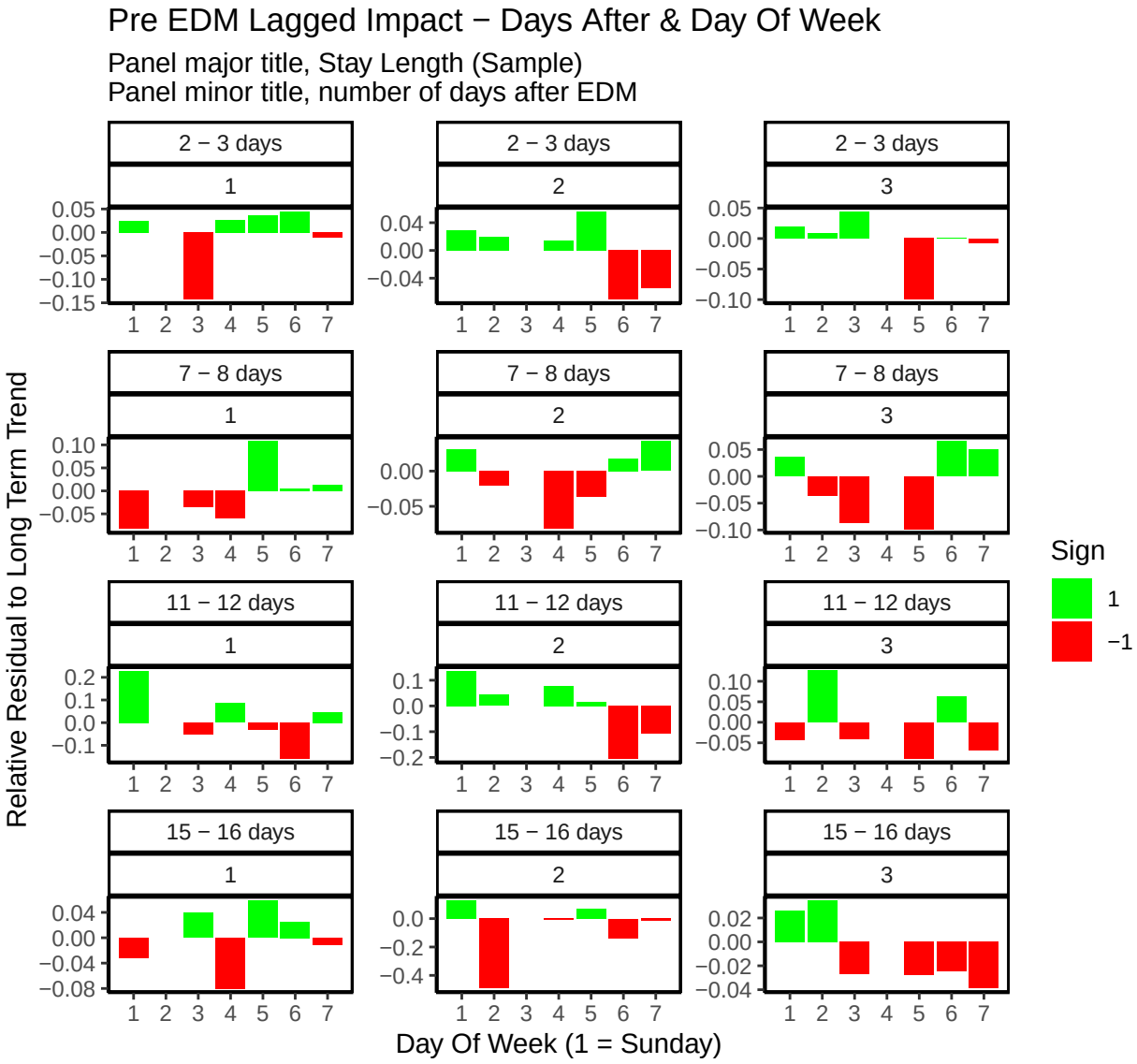
In the chart below, the first column represents non limited time offer EDMs while the second shows limited time offer EDMs. The x-axis represents the ID associated to each stay length. While the IDs are not strictly in order of occupancy time, it tries to represent the varying impact across different stay lengths.



10 - EDM Major Lagged Days (DOW)

Uses Google Analytics and related sales data (post October 2018).

This component further stratifies the previous factor by the day of week of each day after an EDM is sent out.

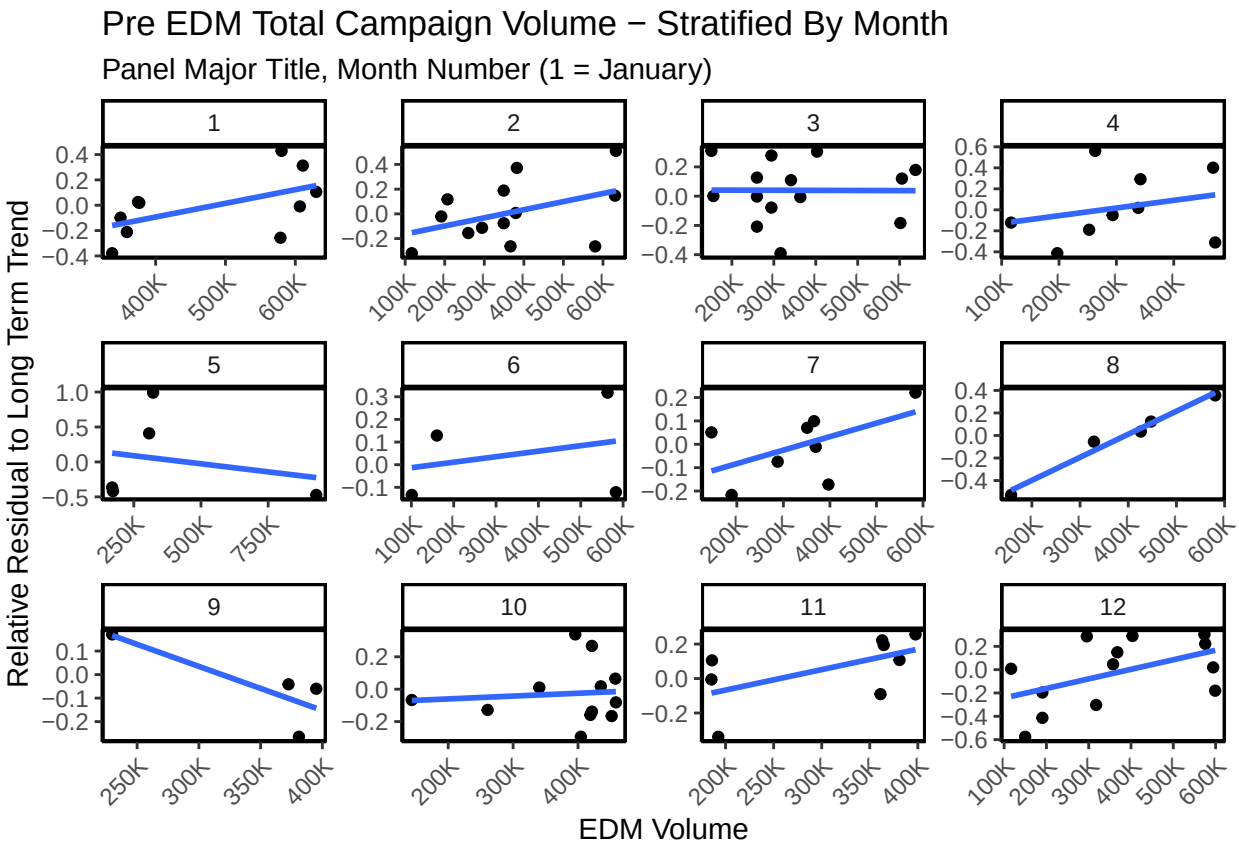


11 - EDM Major Month Regression

Uses Google Analytics and related sales data (post October 2018).

This factor represents the impact of the EDM recipient volume on the residual variability, stratified by month.

In general, it seems that the more recipients are targeted by an EDM, the better the impact on overall sales volume. With the clear exception of September, this behaviour holds true for most months. There seems to be some opportunity to further capitalise on high volume EDMs in July, August, November and December as these months have the steepest slope.

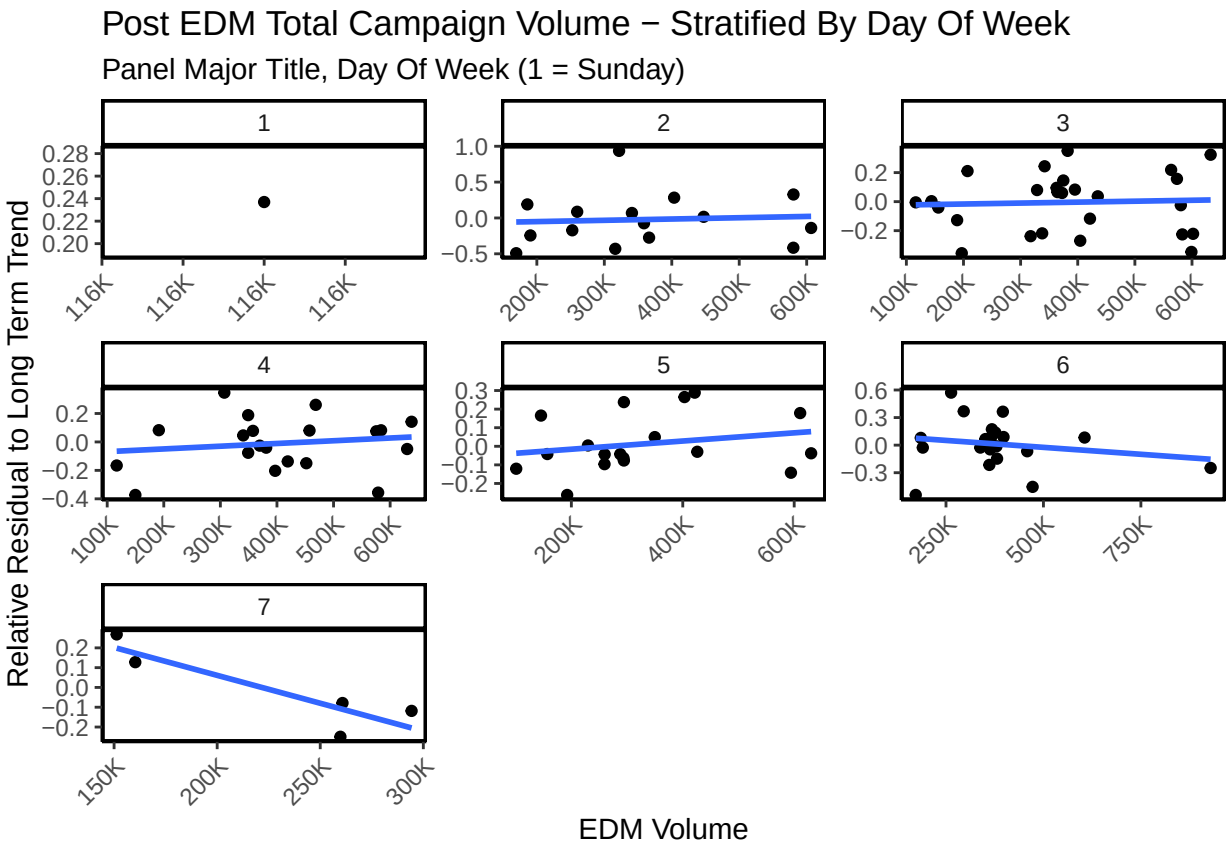


12 - EDM Major Day Of Week Regression

Uses Google Analytics and related sales data (post October 2018).

Similar to the previous component, this factor represents the impact of recipient volume when stratified by the day of week.

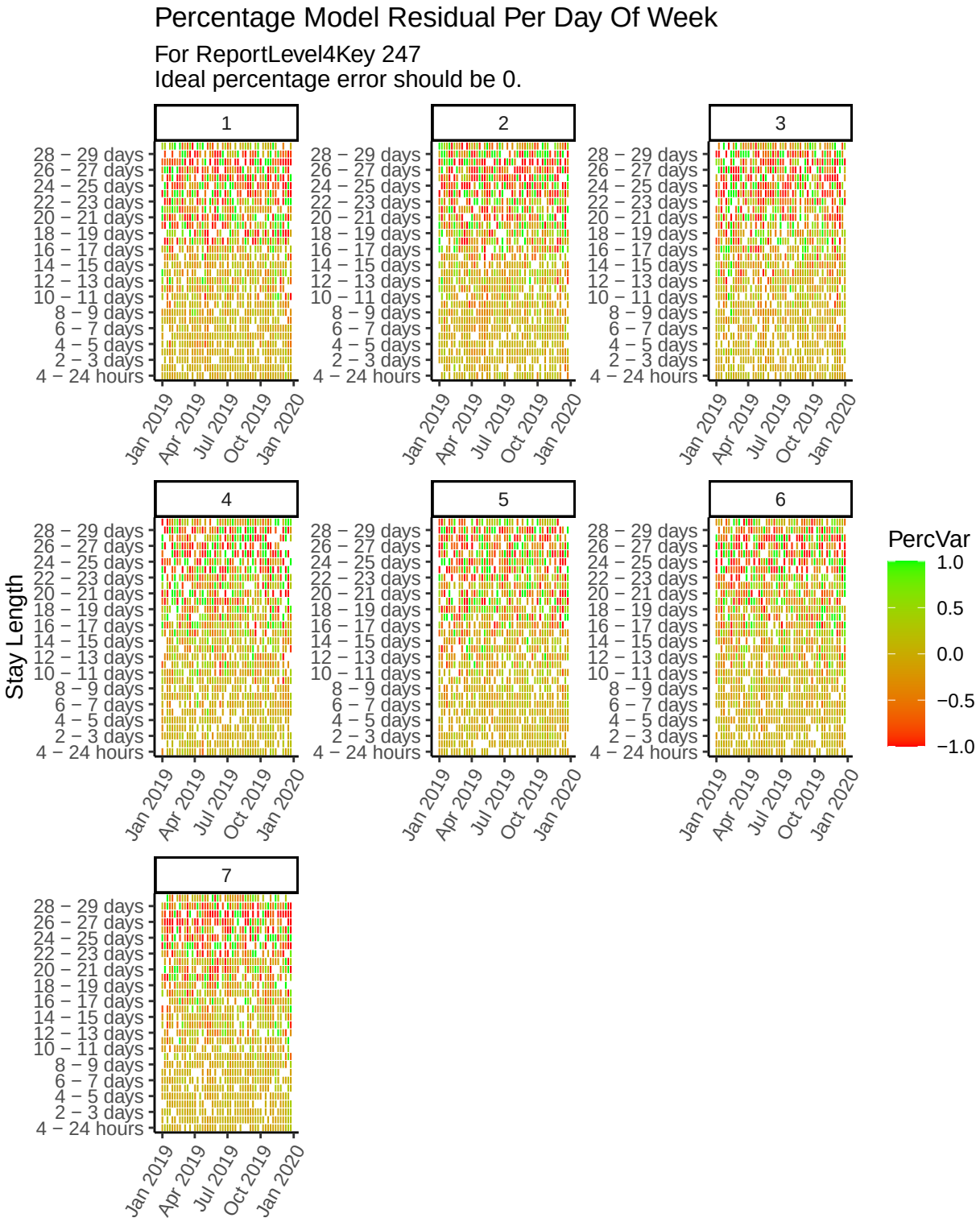
The results are not nearly as significant as the monthly stratification but there is still some impact across the middle of the week.



Remaining Residuals - Sample Size

After accounting for the long term trend approximation and numerous short term factors, the model is left with a relatively low level of residuals. A sample is shown below for ReportLevel4Key 247 and is faceted by day of week for clarity.

For demonstration purposes, the percentage error shown has a floor and ceiling of -1 and 1 respectively.



An initial observation shows that the shorter half of the stay lengths have much lower percentage residuals. In fact, if the floor and ceiling restrictions were removed the longer stay lengths would show more severe gradient extremes, suggesting poor model performance. This is a result of lower volume data population. Commercialised car parks are normally designed to facilitate a continuous flow of customers going in and out. This is especially true for car parks within a business or retail hub where customers generally stay for

shorter periods of time to fulfill work or purchasing commitments. Consequently, there would be significantly lower number of customers who choose to stay longer as there are not many associated commitments that require that length of stay.

This issue can be observed in the below table where the relative size of sales volume, MAE and MBE are calculated for the available dates.

StayLengthName	MAE (%)	MBE (%)	TotalPurchases	%
4 - 24 hours	12	0	33084	11.99
1 - 2 days	10	0	52053	18.87
2 - 3 days	11	0	63417	22.99
3 - 4 days	14	-1	41936	15.20
4 - 5 days	13	-2	27016	9.79
5 - 6 days	15	-3	13088	4.74
6 - 7 days	19	-3	8541	3.10
7 - 8 days	18	-2	13191	4.78
8 - 9 days	20	-3	5325	1.93
9 - 10 days	24	-1	3870	1.40
10 - 11 days	29	0	2591	0.94
11 - 12 days	33	-2	1805	0.65
12 - 13 days	35	-2	1310	0.47
13 - 14 days	35	-1	1245	0.45
14 - 15 days	28	-2	2425	0.88
15 - 16 days	39	-3	1260	0.46
16 - 17 days	47	-2	648	0.23
17 - 18 days	52	-4	442	0.16
18 - 19 days	52	-2	412	0.15
19 - 20 days	62	-4	301	0.11
20 - 21 days	68	-14	239	0.09
21 - 22 days	49	-6	432	0.16
22 - 23 days	64	-7	190	0.07
23 - 24 days	77	-24	128	0.05
24 - 25 days	75	-21	98	0.04
25 - 26 days	76	-24	93	0.03
26 - 27 days	74	-33	63	0.02
27 - 28 days	81	-31	76	0.03
28 - 29 days	73	-18	96	0.03
29+ days	50	3	489	0.18

In addition, the purchasing behaviour of longer stay customers are more sporadic. For example, it is reasonable to assume that customers go on a long holiday once per year which would require the purchase of longer occupancy. This effectively produces a burst in sales volume once a year while the rest of the period receives close to zero. The long term trend approximation assumes that bursts in sales are somewhat carried throughout the averaging window as it tries to smooth the peaks and troughs. In reality a more accurate approximation should partition the time series into periods of high sales and low sales volume. This would allow the short term factors to be calculated from a more realistic trend. As the overall sales volume in longer stay lengths are significantly lower, the general impact when assessing from a ReportLevel4Key perspective is quite low.

Remaining Residuals - Price

While most of the short term variability has been modelled with time dependent and marketing related factors, it is also reasonable to factor in changes in product pricing. Such a relationship would generally fall under a demand and supply model where differing levels of price elasticity are introduced. The marketing related short term factors partially addresses price adjustments as most EDMs come with some discount level. However the component of the EDM that is independently associated with price elasticity is difficult to assess as its purpose is also to communicate and showcase the available products to customers. This can easily be seen in the regression components of the short term factors where EDM recipient volume seems to play a significant role in sales volume performance.

To try and explore the impact of price on sales volume, correlations have been calculated for each stay length between model. These correlations are against:

1. Percentage daily ATV (proxy for price) movements relative to prior day.

- Useful for assessing the impact of aggressive pricing strategies based on prior day prices.

2. Percentage daily ATV movements relative to long term ATV trends.

- Useful for assessing the impact of price adjustments relative to what has been the trend.

Below table shows a sample of the correlation results for ReportLevel4Key 247. Note that "Against Trend" refers to percentage ATV changes relative to its trend while "Against Prior Day" is a day on day percentage change. Unfortunately it is difficult to produce the correlation on an aggregated level as the price points and general sales volume for each product can vary significantly.

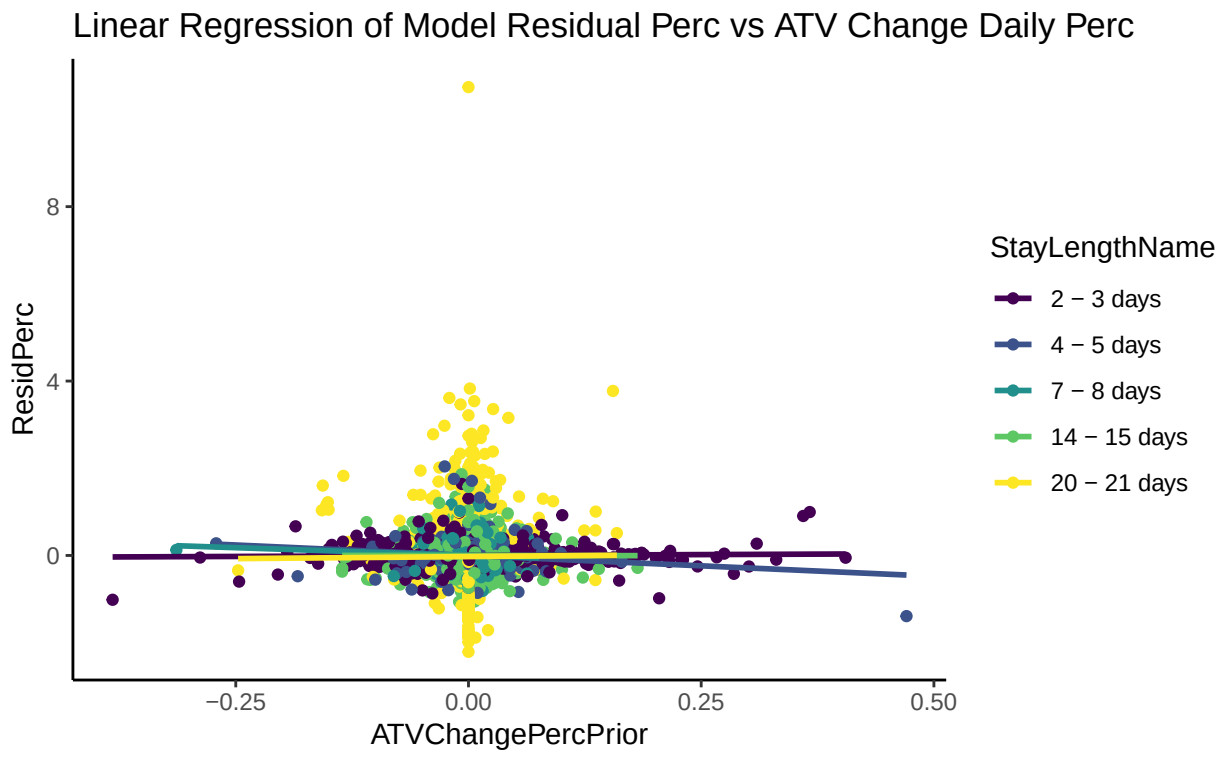
StayLengthName	Against Trend	Against Prior Day	TotalPurchases
15 - 30 mins	0.10	0.13	37
30 - 60 mins	-0.01	0.04	79
1 - 2 hours	0.02	0.06	201
2 - 3 hours	-0.07	-0.02	202
3 - 4 hours	-0.03	0.03	186
4 - 24 hours	-0.01	-0.03	47503
1 - 2 days	-0.02	-0.02	77593
2 - 3 days	0.03	0.03	93016
3 - 4 days	0.08	0.07	62089
4 - 5 days	-0.11	-0.12	40669
5 - 6 days	0.03	0.03	19836
6 - 7 days	-0.01	-0.01	13075
7 - 8 days	-0.06	-0.06	22354
8 - 9 days	0.00	0.00	8377
9 - 10 days	-0.15	-0.13	5822
10 - 11 days	-0.11	-0.11	3945
11 - 12 days	0.07	0.08	2828
12 - 13 days	-0.01	-0.01	2058
13 - 14 days	0.05	0.05	1924
14 - 15 days	0.00	0.00	4155
15 - 16 days	-0.19	-0.17	2061
16 - 17 days	-0.04	-0.02	1010
17 - 18 days	0.02	0.03	709
18 - 19 days	0.01	0.01	656
19 - 20 days	0.01	0.02	459
20 - 21 days	0.00	0.01	391
21 - 22 days	0.03	0.03	774

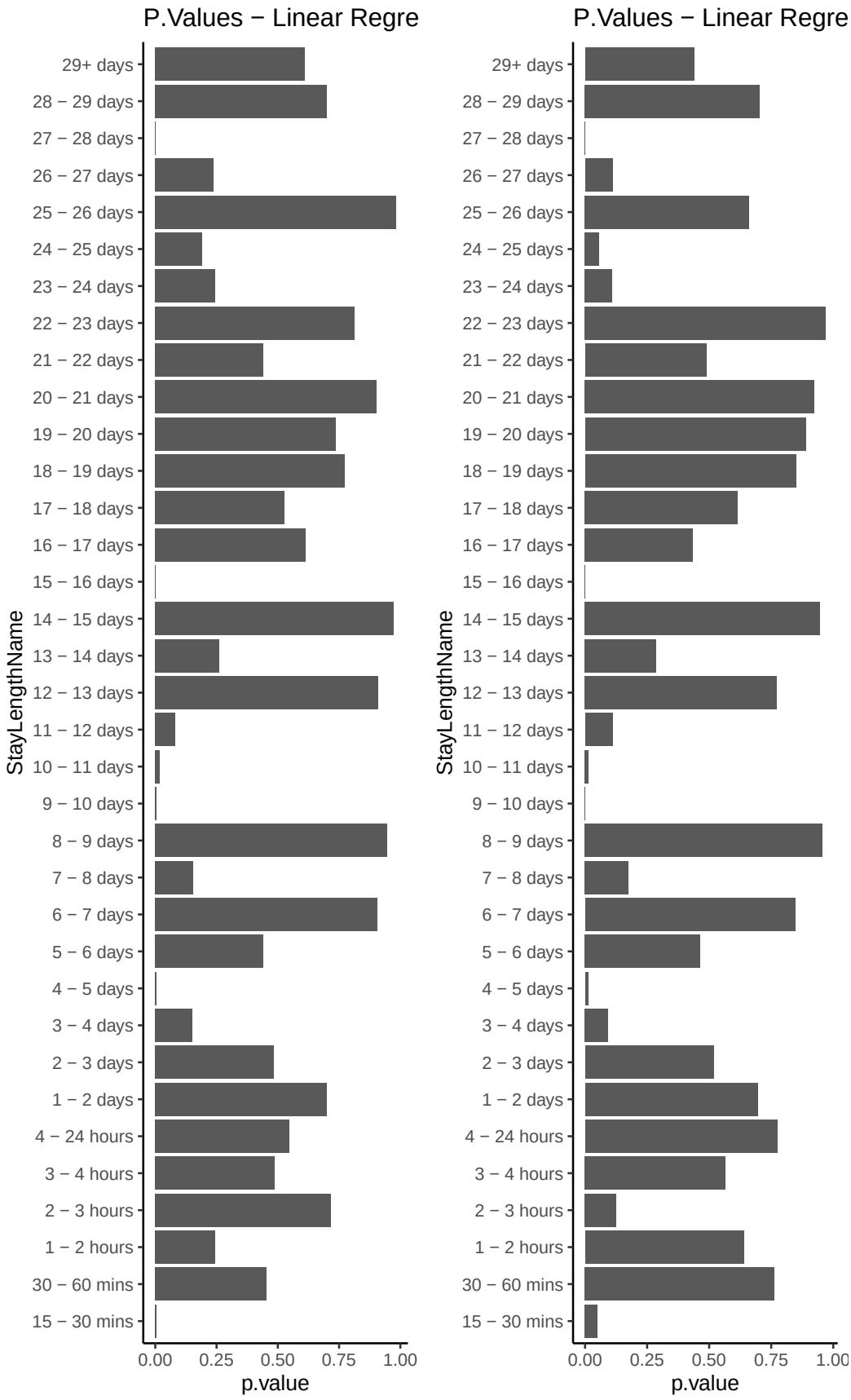
StayLengthName	Against Trend	Against Prior Day	TotalPurchases
22 - 23 days	0.00	0.01	335
23 - 24 days	-0.07	-0.05	213
24 - 25 days	-0.09	-0.06	165
25 - 26 days	-0.02	0.00	139
26 - 27 days	-0.07	-0.05	108
27 - 28 days	-0.16	-0.14	144
28 - 29 days	-0.02	0.02	168
29+ days	-0.04	-0.02	732

In general, the final model residuals show weak correlation, whether positive or negative, to price movements. There are some stay lengths that show better signs, such as 4 - 5 days, however these correlations are still low.

To explore a different view, the model percentage residuals have also been regressed against the percentage daily price movements. The p.values for each stay length are shown in the below charts and shows a similar story to the correlations. Most stay lengths are well above a .05 significance level, making it difficult to justify using price as a reasonable explanatory variable to the remaining residual. However some stay lengths are statistically significant, such as the 4 - 5 day band. While there is potential to use price on statistically significant stay lengths, the relative explanatory power is quite low, as can be seen in the slopes generated below.

A balance of explanatory power, in terms of regression coefficient, and achieving statistically significance are required to justify the use of ATV change to explain the remaining residuals.




P.Values − Linear Regre
P.Values − Linear Regre
StayLengthName
29+ days
28 − 29 days
27 − 28 days
26 − 27 days
25 − 26 days
24 − 25 days
23 − 24 days
22 − 23 days
21 − 22 days
20 − 21 days
19 − 20 days
18 − 19 days
17 − 18 days
16 − 17 days
15 − 16 days
14 − 15 days
13 − 14 days
12 − 13 days
11 − 12 days
10 − 11 days
9 − 10 days
8 − 9 days
7 − 8 days
6 − 7 days
5 − 6 days
4 − 5 days
3 − 4 days
2 − 3 days
1 − 2 days
4 − 24 hours
3 − 4 hours
2 − 3 hours
1 − 2 hours
30 − 60 mins
15 − 30 mins
0.00
0.25
0.50
0.75
1.00
p.value

All Factors

The model produces twelve different factors that influence the short term variability in online parking sales volume, plus the approximation for the long term trend. These can be categorised into logical groups as below.

DecompositionType_l1	DecompositionType_l2	DecompositionName
Model Results	Long Term Trend	Predict_TrendLong
Model Results	Seasonality	Predict_SeasonalMonth
Model Results	Seasonality	Predict_SeasonalDOW
Model Results	Seasonality	Predict_SeasonalSchoolHoliday
Model Results	Seasonality	Predict_SeasonalInfluentialDate
Model Results	EDM Initial Impact	Predict_EDMMajorAvg
Model Results	EDM Initial Impact	Predict_EDMMajorMonth
Model Results	EDM Initial Impact	Predict_EDMMajorDOW
Model Results	EDM Lagged Impact	Predict_EDMMajorLaggedAvg
Model Results	EDM Lagged Impact	Predict_EDMMajorLaggedDays
Model Results	EDM Lagged Impact	Predict_EDMMajorLaggedDaysDOW
Model Results	EDM Initial Impact	Predict_EDMMajorMonthRgr
Model Results	EDM Initial Impact	Predict_EDMMajorDOWRgr

These groupings are designed to bridge the technical calculations and common business understanding. Prior to undertaking this analysis, the daily sales volume performance have been usually associated to seasonality and marketing related factors. As such the categorisation produced in this analysis naturally links with current business understanding and will help in future decision making.

Training Data Format

Each short term factor sensitivity is stored in separate tables. The relevant grouping dimensions are also included so that they can be re-integrated back into the training, validation or a similarly formatted data frame.

Below is a sample of what the training sensitivity data table looks like.

ReportLevel4Key	StayLengthKey	DimAmountTypeKey	MonthNo	TrendLongRatio
247	10	103	1	-0.0098471
247	10	103	2	0.0316058
247	10	103	3	0.0076535
247	10	103	4	-0.0431604
247	10	103	5	0.0749755
247	10	103	6	-0.1089029
247	10	103	7	-0.0166079
247	10	103	8	0.0683657
247	10	103	9	0.0596189
247	10	103	10	0.2305259
247	10	103	11	0.0907740
247	10	103	12	-0.2654299

Model Accuracy - Training & Test Sets

Training Accuracy

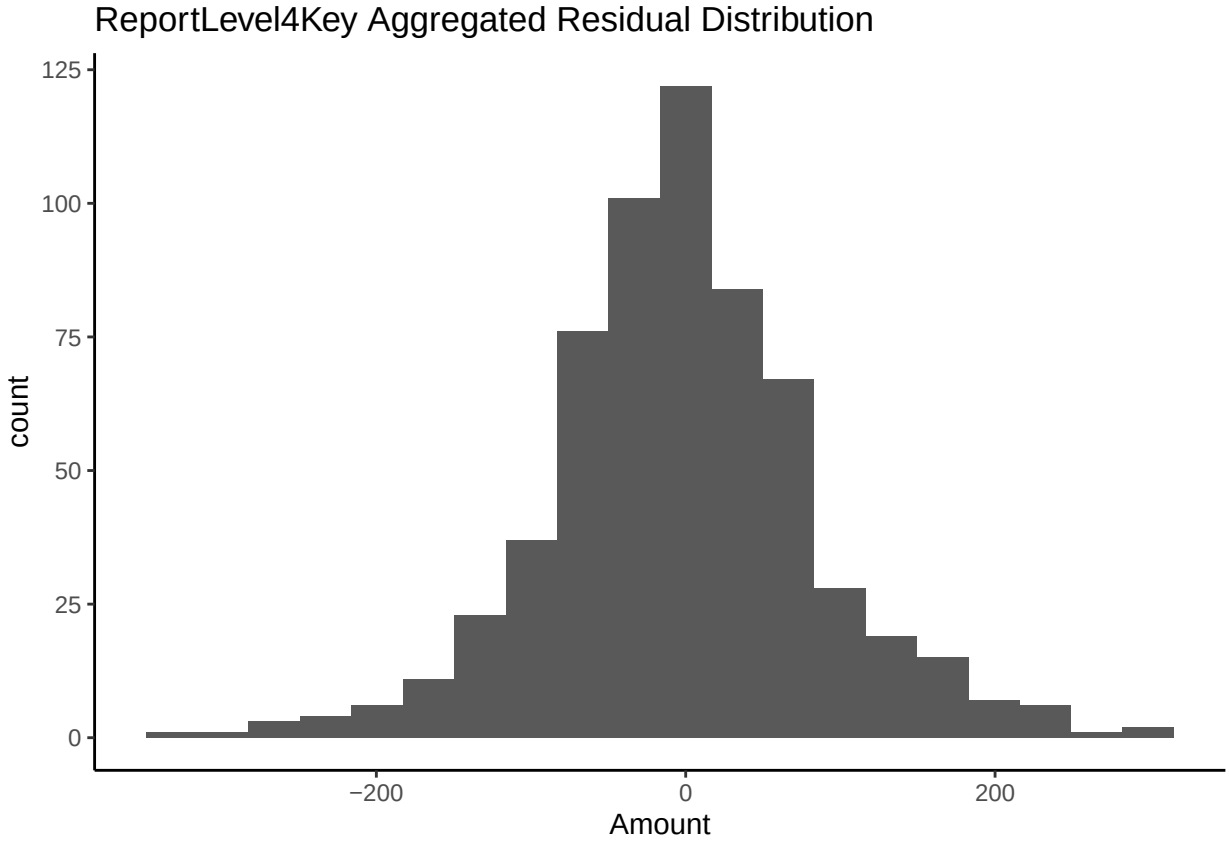
Using the function "applyTrainingResults", the following summary accuracy measures have been produced for both training and test validation data. The relative performance of the model should be evaluated per ReportLevel4Key and StayLengthName as this is the common level of granularity for each short term factor. However there are too many groupings to cover in this report alone and a higher level summary will be used instead.

Data	MAE	MBE	RMSE	MAEPerc	MBEPerc	RMSEPerc
Test	46.74812	-1.680146	68.24075	0.1992678	-0.0770721	0.1882808
Train	120.33353	-1.212209	163.88573	0.1730367	-0.0815931	0.1178015

It is also important to note the reason that the RMSE Perc result is lower than the MAE Perc, despite the opposite for the raw values, is because it is calculated by dividing the final RMSE by the group average sales volume. It washes out a lot of the finer variability that comes with the per day or per stay length granularity. Drilling down to stay length stratification brings the two percentage measures closer to each other. An example of this can be seen below where the error rate measures are shown per stay length of ReportLevel4Key 247 in the training data.

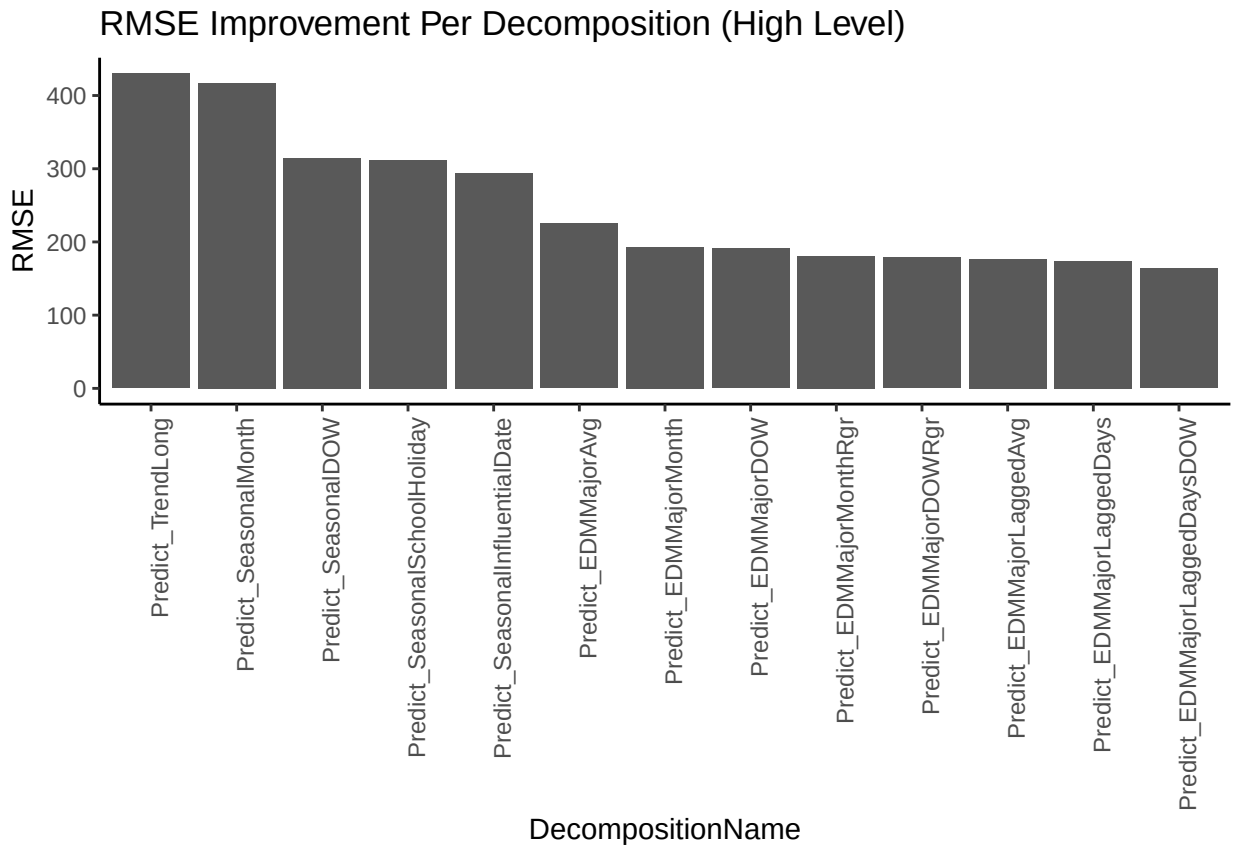
StayLengthName	MAE	MBE	RMSE	MAEPerc	MBEPerc	RMSEPerc
4 - 24 hours	13.34	0.07	18.33	0.24	-0.04	0.19
1 - 2 days	18.53	-0.65	25.61	0.23	-0.06	0.16
2 - 3 days	24.44	-1.24	32.78	0.26	-0.10	0.17
3 - 4 days	18.33	-0.47	24.25	0.25	-0.11	0.19
4 - 5 days	12.05	-0.48	17.01	0.29	-0.14	0.21
5 - 6 days	6.41	-0.31	8.36	0.21	-0.06	0.21
6 - 7 days	5.44	-0.33	7.15	0.27	-0.11	0.27
7 - 8 days	9.23	-0.07	12.30	0.33	-0.18	0.27
8 - 9 days	3.74	-0.15	5.04	0.30	-0.12	0.30
9 - 10 days	2.97	-0.11	3.93	0.33	-0.13	0.33
10 - 11 days	2.56	0.06	3.45	0.47	-0.24	0.41
11 - 12 days	1.95	0.06	2.49	0.45	-0.19	0.43
12 - 13 days	1.60	0.07	2.03	0.52	-0.23	0.48
13 - 14 days	1.48	0.08	1.96	0.50	-0.24	0.49
14 - 15 days	2.48	0.04	3.32	0.43	-0.22	0.38
15 - 16 days	1.74	0.08	2.24	0.61	-0.31	0.54
16 - 17 days	1.03	0.32	1.39	0.47	-0.08	0.69
17 - 18 days	0.88	0.36	1.14	0.46	0.00	0.79
18 - 19 days	0.88	0.44	1.26	0.44	0.05	0.94
19 - 20 days	0.71	0.50	0.88	0.40	0.21	0.93
20 - 21 days	0.71	0.53	0.83	0.42	0.27	1.03
21 - 22 days	0.92	0.34	1.23	0.47	-0.02	0.79
22 - 23 days	0.72	0.59	0.75	0.46	0.35	1.12
23 - 24 days	0.70	0.65	0.63	0.51	0.46	1.44
24 - 25 days	0.73	0.69	0.56	0.58	0.54	1.64
25 - 26 days	0.70	0.66	0.48	0.57	0.53	1.73
26 - 27 days	0.70	0.69	0.42	0.61	0.60	1.91
27 - 28 days	0.70	0.67	0.49	0.57	0.54	1.69
28 - 29 days	0.82	0.81	0.58	0.61	0.60	1.76
29+ days	0.90	0.38	1.20	0.44	0.01	0.78

Using MAEPerc and RMSEPerc as measures shows 17% and 12%, respectively, for the high level training data error rate. The MBE is close to zero which suggests that the model does not suffer from general under or over prediction. However the MBE Perc is not close to zero but can be explained by the omission of NA, Inf and -Inf results. Below visualisation shows the residuals for ReportLevel4Key 247 which is normally distributed around 0, supporting the low bias suggested from the MBE.



As covered in the residual section, the relative error of the predictions increase as we get to longer stay lengths. This seems to be due to data population and long term trend calculation issues. As such it would be reasonable to say that the accuracy of this model is better suited for the shorter stay products where sales volume is more consistent throughout the year.

While overall error is an important metric, it is also good to understand where most of the improvements comes from. This can be observed by plotting out the RMSE improvement per additional short term factor. The most significant improvements come from the SeasonalDOW, EDMMajorAvg and EDMMajorMonth factors while the remaining shows smaller contribution. As the predictive contribution of individual factors depend on the order they are calculated, it would be worthwhile to average out their contribution across different ordering. Nevertheless, the combined power of each factor significantly improves the predictive power from just using the long term trend as the data shows strong time dependent and marketing variability.



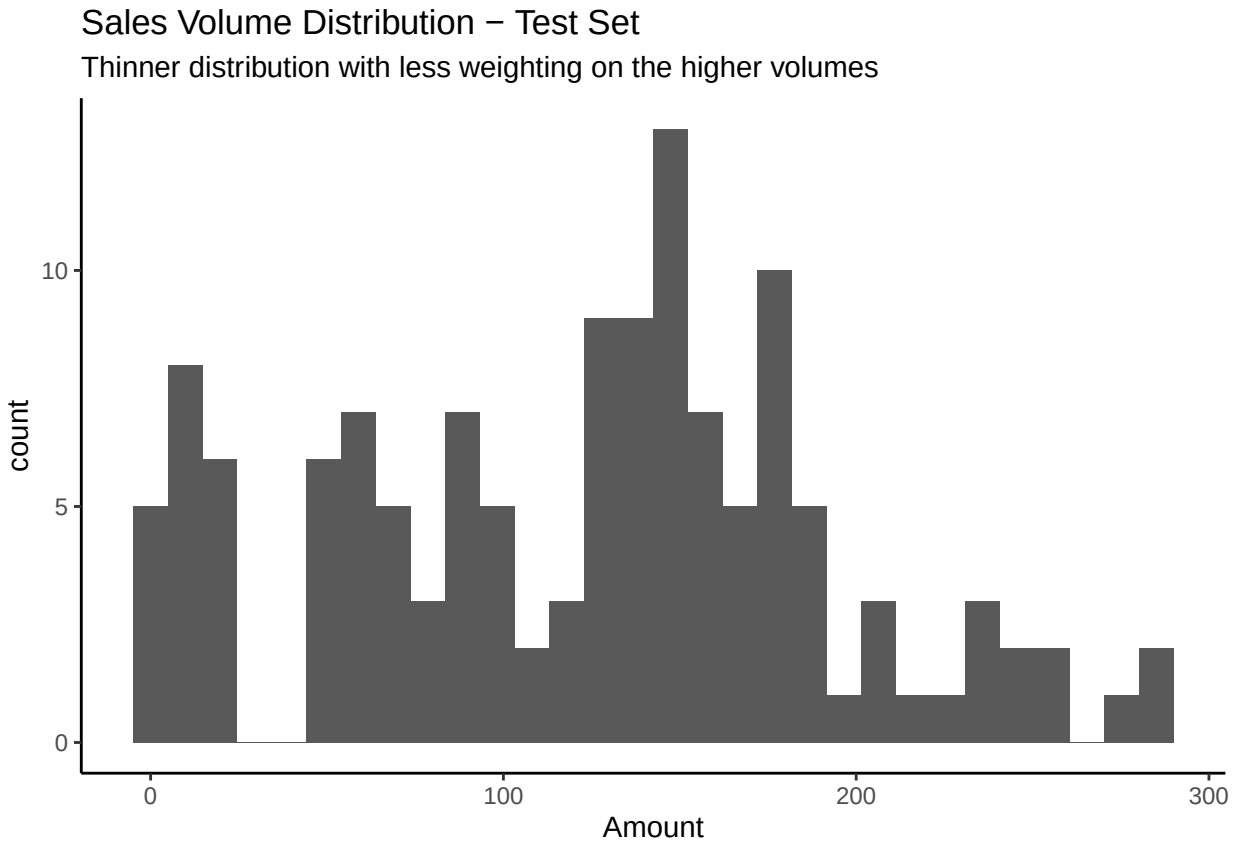
DecompositionName	RMSE	% Change
Predict_TrendLong	430.1108	NA
Predict_SeasonalMonth	416.7305	-3
Predict_SeasonalDOW	315.1258	-24
Predict_SeasonalSchoolHoliday	311.7063	-1
Predict_SeasonalInfluentialDate	293.9041	-6
Predict_EDMMajorAvg	225.1979	-23
Predict_EDMMajorMonth	192.6984	-14
Predict_EDMMajorDOW	190.9344	-1
Predict_EDMMajorMonthRgr	180.8908	-5
Predict_EDMMajorDOWRgr	178.9326	-1
Predict_EDMMajorLaggedAvg	176.4985	-1
Predict_EDMMajorLaggedDays	174.1661	-1
Predict_EDMMajorLaggedDaysDOW	163.8857	-6

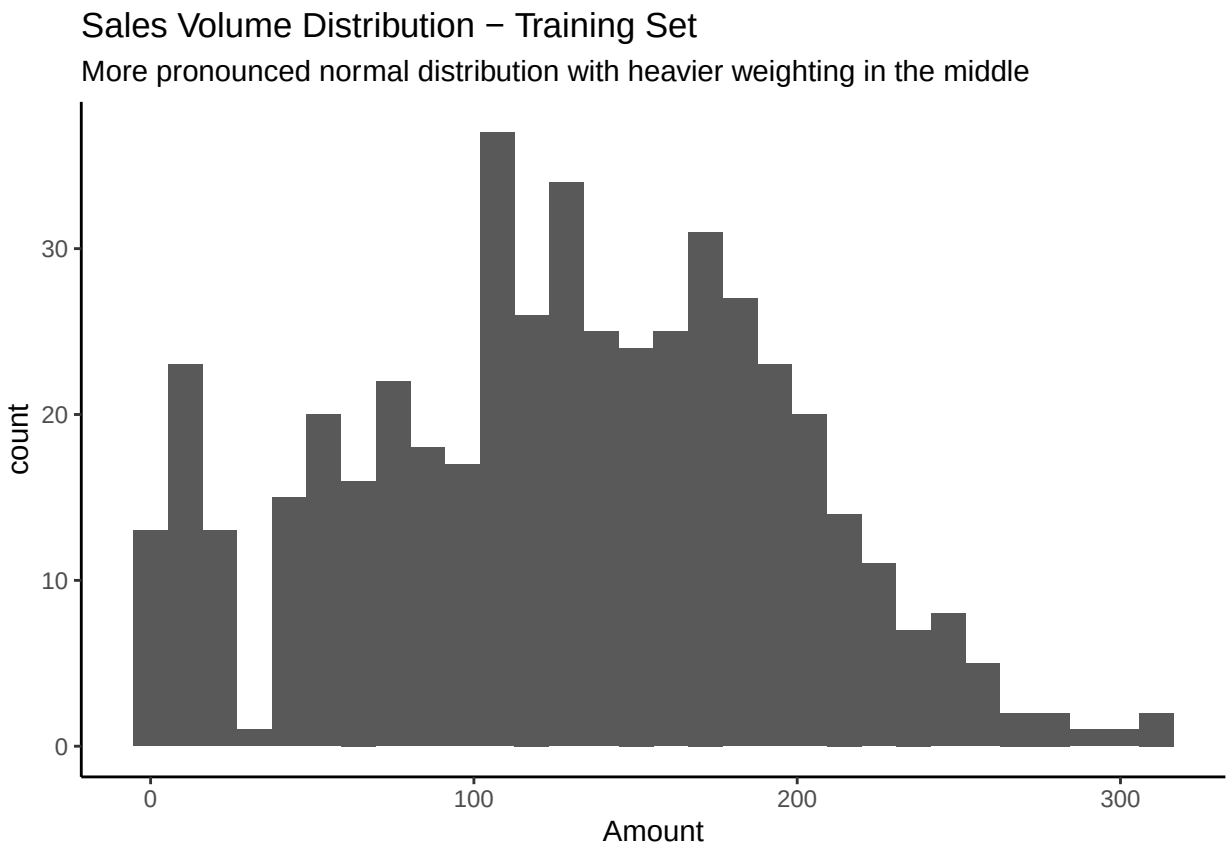
Test Validation Accuracy

Data	MAE	MBE	RMSE	MAEPerc	MBEPerc	RMSEPerc
Test	46.74812	-1.680146	68.24075	0.1992678	-0.0770721	0.1882808
Train	120.33353	-1.212209	163.88573	0.1730367	-0.0815931	0.1178015

The raw MAE and RMSE values for the test set is significantly lower than the training set. However this does not tell the full story as there are other factors at play. To better gauge test set versus training set difference, it is better to look at the MAEPerc and RMSEPerc. They are both consistently higher than their traning dataset counterparts which is to be expected. MBE measure is also close to zero which suggests that the model is generally unbiased when applied to unseen data. While generally higher than the training set, the error rate metrics are still below 20% which is a good result considering this model is meant to be used on a daily basis.

The significant difference between the raw MAE and RMSE values can be attributed to the data partitioning process and the granularity of the data. The partitioning was performed on a total dataset level while the training is performed on a product level. Consequently, the partition may unevenly distribute products with higher sales volume on some dates towards the training set on a per product level as the 80/20 split is performed on a higher level of granularity. This results in the partition samples becoming biased on a total sales volume perspective which directly impacts the MAE and RMSE results. The partitioning issue can be easily seen in the distribution plots below.





To avoid this issue, the partitioning should be done on a per product level to match the training algorithm.

Overtraining

As covered earlier, the marketing related factors suffer from a smaller volume of data due to a shorter date range. Consequently, the different levels of groupings in the EDM factors can lead to over training as the sample size can be quite small. While this is a real issue, the impact on the accuracy measures when moving from training to test data sets is not significant enough to suggest over training. The operational use of this model lends itself to be retrained periodically with larger datasets. This will naturally get to a point where the group sizes per factor will get large enough such that the data vailability issue will become negligible.

Challenges & Potential Improvements

While the report demonstrates a clear outline to decomposing the short term variability, there have been plenty of challenges and room for improvement.

- Data partitioning
  - Refer to "Test Validation Accuracy" section for more details.
  - Grouped K Folds within "createDataPartition" should be used in future.
- Error rate
  - While the sub 20% error rate is reasonable there is potential for further improvement by trying to factor in price adjustments (differently to what was explored earlier) and website performance as reported in Google Analytics.

  - Applying algorithms within the caret suite to further reduce the error rate.
    * The main caveat of using other algorithms is their ability to translate to common business understanding.

- Long Term Trend
  - While the focus is on short term analysis, the long term trend still plays an important role as it allows the factors to be calculated on a relative level.
  - More sophisticated methods than centered moving average should be used as it has difficulty handling seasonal bursts of sales activity (covered in " Remaining Residuals - Sample Size").

- Matching data availability
  - The only way to fix this is to somehow backdate Google Analytics data to match the date range of the sales volume.

- Sampling and over training
  - Refer to the previous section “Overtraining”.

- Decomposition order
  - As explored in “Short Term Trend - Overview”, the order of decomposition influences the individual impact of each factor. Averaging out the resulting impact from all ordering combinations may result in a balanced view of each factor.

As mentioned previously, the intent behind this report is for operational business use. In doing so, constant improvements, iterations and more data will generally alleviate a lot of the issues listed above.

Summary

In summary the model has set successfully decomposed a large proportion of the daily variation experienced within the online parking sales data. The factors produced from the model are easily translatable to common business understanding which easily allows the model to be built up. The model itself is also quite dynamic and easily allows for new factors to be plugged in. In addition, the model also provides a framework to build quantitative factors that easily support future business decisions. As outlined in the previous section, there many avenues for improvement which will be actively explored in the coming months to further increase the accuracy of the model.