The Alan Turing Institute

Data Study Group Final Report: ASDA

12 – 23 July 2021

Exploring and quantifying the effect of weather on sales

https://doi.org/10.5281/zenodo.6498350

Contents

1 Executive summary

1.1 Challenge overview

ASDA, one of the biggest supermarket chains in the UK, is interested in how weather affects sales of certain product groups. Understanding the weather-sales relationship would allow ASDA to manage the supply-chain system and distribution in a timely and efficient manner, particularly reducing the risk of food waste for fresh products such as meat while having products in stock as needed. The analysis of the weather-sales relationship builds on a huge dataset which includes daily sales data for 151 product groups in over 600 stores across the UK over 3 years, and daily weather conditions at each store, including mean daily humidity, wind speed, rainfall, snowfall and minimum and maximum daily temperature.

The Alan Turing Institute is the UK's national institute for data science and artificial intelligence, with headquarters at the British Library. Data Study Groups are intensive five day collaborative hackathons hosted at the The Alan Turing Institute, which bring together organisations from industry, government, and the third sector, with multi-disciplinary researchers from academia. ASDA, the Data Study Group Challenge Owner, provided the real-world challenge and the data to be tackled by a group of researchers led by two Principal Investigators and a Facilitator. This report is the culmination of that process and is the result of their joint co-authorship.

1.2 Data overview

We worked with the following dataset entirely provided by the Challenge Owner:

Weather data: historical time series of mean daily humidity, wind speed, rainfall, snowfall and minimum and maximum daily temperature over the period 1/1/2017-29/2/2020 at each ASDA store.

Sales: historical time series of daily sold quantities for 151 selected Product Profile Groups (PPGs) across approximately 60 departments (ODs) at over 620 ASDA stores over the period

1/1/2017-29/2/2020.

Store information: store type and other pieces of information (e.g., approximate location) about each store.

Additionally, we compiled a list of festivities and bank holidays over the period 1/1/2017-29/2/2020 to account for the fact that these events may have an effect on the sales which may confound the sale dependence from the weather.

1.3 Main objectives

The overall goal of this DSG was to understand if and how weather affects specific product sales. The DSG participants addressed the following research questions:

1. Which weather conditions (defined as statistics of a specific weather variable or combination of variables) affect sales? Of which product profile groups?
2. Are the relationships between weather and sales different across stores?
3. Are relationships between weather and sales different as we approach different festivities and across different geographical areas?

1.4 Approach

The participants performed first an extensive exploratory analysis of the data, which highlighted several important costumer buying patterns and identified the variables that have a stronger relationship with the sales. This informed the following phase where the participants applied a range of different statistical and machine learning methods to the dataset to predict sale patterns. The statistical methods included linear models, generalised linear models and multilevel regression while the machine learning methods included a range of random forest methods.

This two-step approach was applied to the entire dataset whenever possible, but restricted to few subsets of the data because the analysis on

the entire dataset proved intractable from a computing perspective, given the 2-weeks time duration of the DSG.

1.5 Main conclusions

The participants identified the variables which sho a stronger relationship with sales. These are temperature, day of the week, month, wind and humidity. In particular, it is important to consider anomalies from the long term mean when analysing the weather variables to remove the effect of their natural seasonal fluctuations which may confound the relationship with sales.

The machine learning approaches were extremely fast and useful in gaining an understanding of the relationships between the sales and the other variables but, being *black-box* models, are usually hard, if not impossible, to interpret. Because the interpretability of the model outputs is very important Challenge Owner, we recommend to adopt approaches which merge the best characteristics that both statistics and machine learning have to offer and provides models which provide an inferential understanding of the relationships in the model, but also improved forecasting ability.

1.6 Limitations

The dataset comprised of over 93 million data points and this posed a computational limitation to the exploratory analyses and the experiments because of the 2-weeks duration of the DSG. Therefore, the group decided to focus most of the analyses on subsets of the data only, e.g., randomly-selected product profile groups or stores. We recommend that the results presented in this study are further validated across all the product profile groups and stores using the complete dataset.

1.7 Recommendations and future work

Here we have provided the Challenge Owner with results of a range of methods on subsets of the data, outlining their strengths and limitations. The main recommendation is to carry out these analysis on larger subsets

of the data, or the full dataset if possible, for more robust estimates and predictions.

More broadly, this challenge provides plenty of opportunities for follow-up work, including extending the work to the entire product portfolio at ASDA, refining the sales forecast model and understanding the impact of weather forecast and its uncertainty on the whole supply chain.

The analysis could also benefited from availability of consumer information along with socio-economic and demographic factors and the store geo-location which can be utilized to develop 'spatially-dependent' models with explicit costumer preferences.

2 Data overview

2.1 Dataset description

We worked with a dataset entirely provided by the Challenge Owner which included the pieces of information that the Challenge Owner considered relevant for achieving the challenge goals. The dataset included:

Weather data: historical time series of mean daily humidity, wind speed, rainfall, snowfall and minimum and maximum daily temperature from 1/1/2017 at each ASDA store.

Sales: historical time series of daily sold quantities for 151 selected Product Profile Groups (PPGs) across approximately 60 departments (ODs) at over 620 ASDA stores from 1/1/2017.

Store information: store type and other pieces of information (e.g., approximate location) about each store.

Additionally, we compiled a list of festivities and bank holidays to account for the fact that these events may impact sales, possibly confounding the sale dependence from the weather.

More precisely, the dataset included sales and weather data for all of the 620 ASDA stores in the UK (340 stores in the North and 280 in the South) and 151 unique product profiles sold in these stores. Stores are identified with an integer number between 4126 and 5900. The dataset included

393 superstores, 31 supercentres, 157 supermarkets and 39 stores adjacent to a petrol station. The store location was classified according to the following categories: Edge Of Centre (126), Town Centre (70), Out Of Town (163), Retail Park (44), District Centre (66), Petrol Station (35), Suburban Centre (60), Destination (43), others (13). Stores were further classified depending on whether sales are influenced by the student population (509) or not (111). Note that, for each store, we defined a unique data identification number as OD $\times$ 10,000+PPG because often PPGs are common across multiple ODs. The weather data comprised six weather variables, namely daily minimum and maximum temperature and daily mean humidity, wind speed, precipitation and snow amount. In total, the dataset included roughly 620x6x1200 $\approx 4 \times 10^6$ observations.

2.2 Data quality issues

The snow and rain data included in the dataset were rounded to the nearest integer value. Representing the daily amount in mm, this format was unsuitable to properly represent the inter-daily and inter-seasonal variability. We did not identify any other data quality issues.

3 Data visualisation

We performed an exploratory analysis of the dataset by focusing on subsets of the data, e.g., randomly-selected product profile groups or stores. This section includes some of these analysis as an example.

Figure 1 shows the sales time series of product OD 1 PPG 120 at store 4126. The product and store are chosen arbitrarily to reduce the number of data to be visualized. We only considered the sales data for the time interval between 01-02-2017 and 29-02-2020 to remove the effect of COVID-19 on sales (which was particularly evident from the inception of the pandemic, i.e., from March 2020). First, we looked at the average amount of daily sales by day of the week, by week in the year and by month of the year. Figure 2 shows that sales are lowest on Sundays and highest on Fridays and Saturdays. Given that the specific time series spans three years, the statistics for each day are based on roughly 160

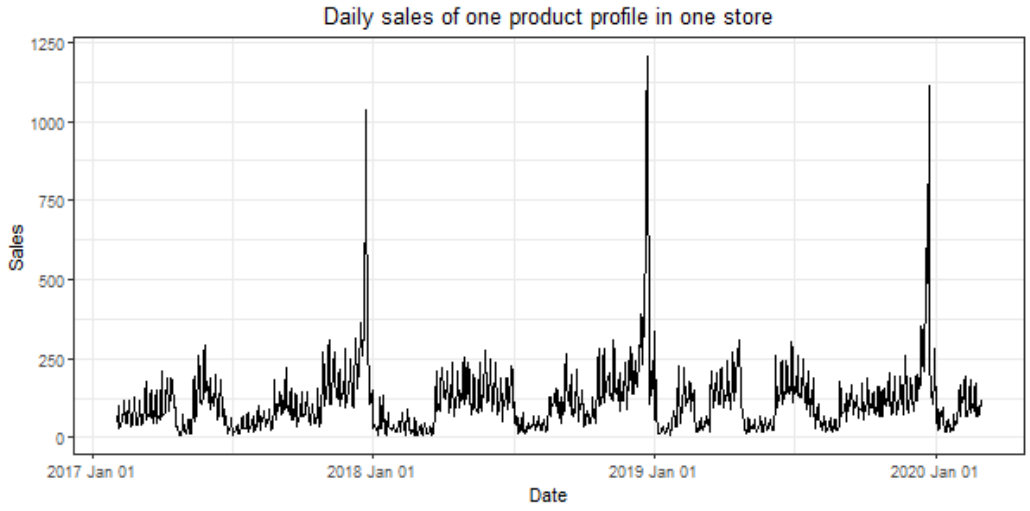

Figure 1: Time series of sales for product OD 1 PPG 120 at store 4126.

observations. The number of outliers exceeding the upper quartile is a minor proportion of all observations (black dots in Figure 2).

Figure 3 shows the average daily sales clustered by week of the year. In the period leading to Christmas (week 51), sales are particularly high.

We also identified different sales patterns during the week. Figure 4 shows high sales on Friday and Saturday and low sales in the rest of the week. The histogram at the top shows the frequency for all daily sales for PPG 120 and OD 1 at store 4126 over the 3 years time interval considered. The other two plots show respectively the sales clustered according to either being on a Friday or Saturday or otherwise being in another day of the week.

Figure 5 shows the time series for each weather variable for store 4126. At a first glance, the evolution over time of humidity, wind and temperature presents relatively high levels of short term variation in their values (e.g. from one day to the next). We hypothesized that such variation can be informative in the investigation of weather-sales relationship. We supposed that consumer behaviour is more likely to deviate from average when weather conditions deviate from average. For each weather

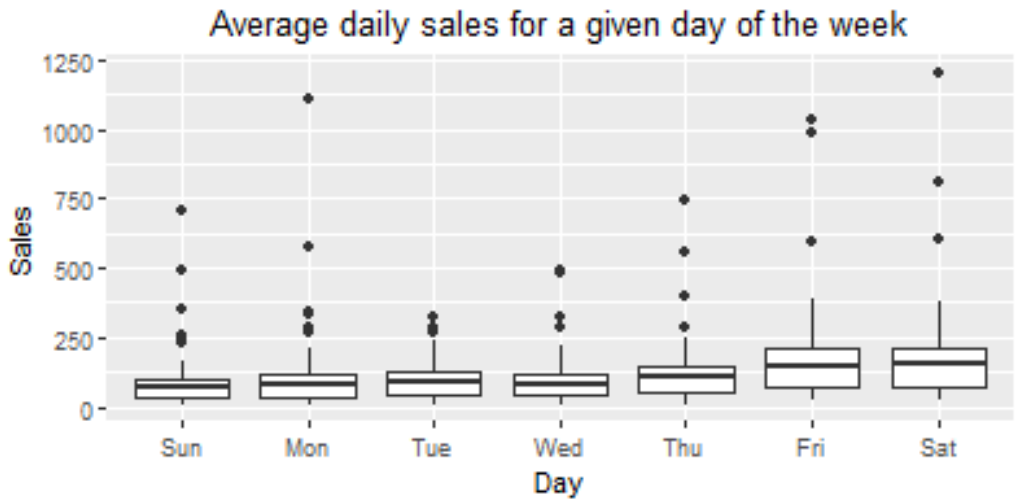

Figure 2: Boxplot of the sales per day of the week for product OD 1 PPG 120 at store 4126.

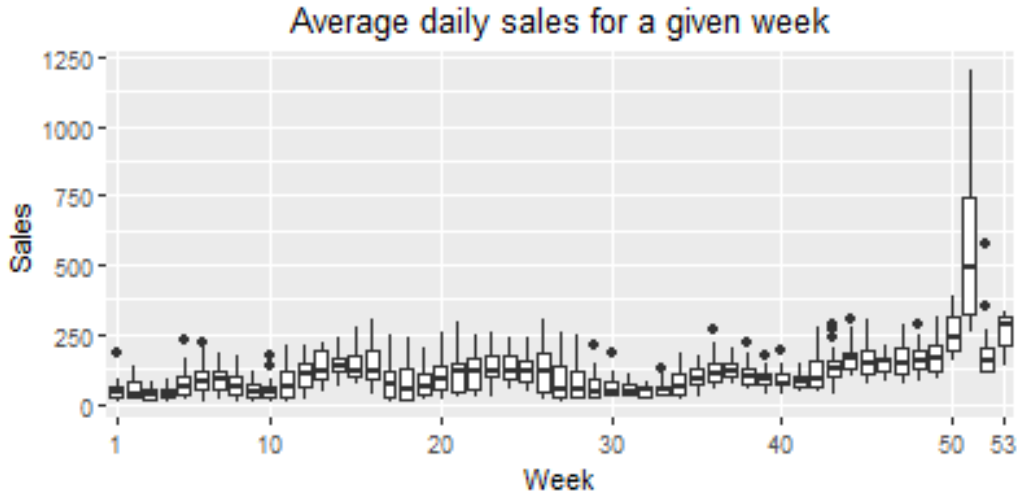

Figure 3: Boxplot of the sales per week of the year for product OD 1 PPG 120 at store 4126.

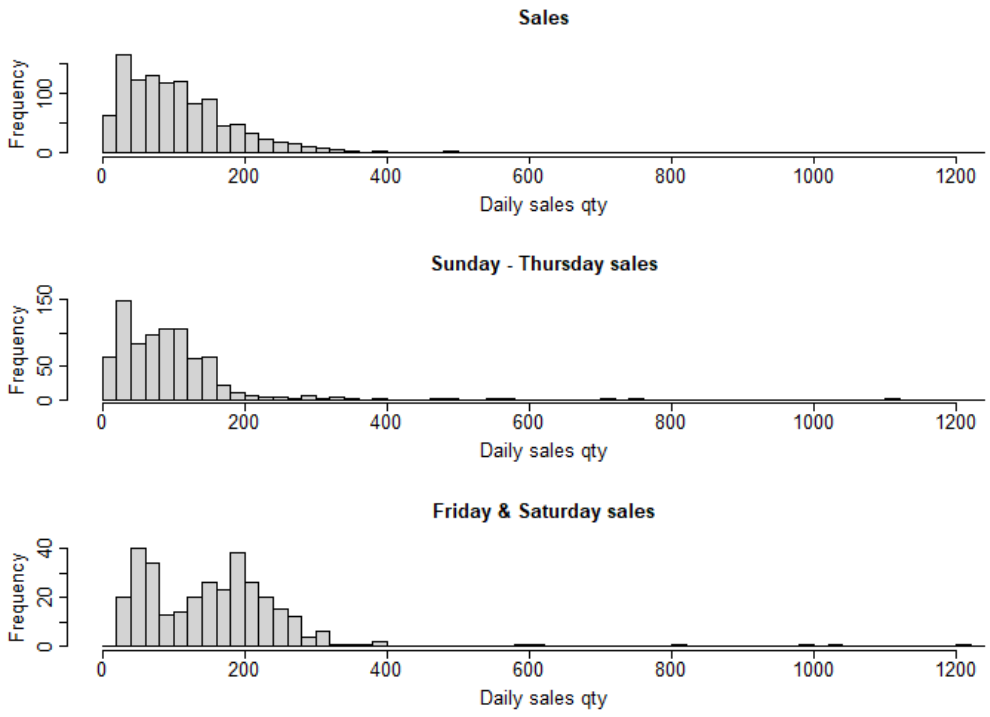

Figure 4: Empirical frequency distribution of sales for PPG 120 OD 1 at store 4126 over the 3 years time interval considered for all days of the week (top), for Friday and Saturday only (bottom) or another day of the week (middle).

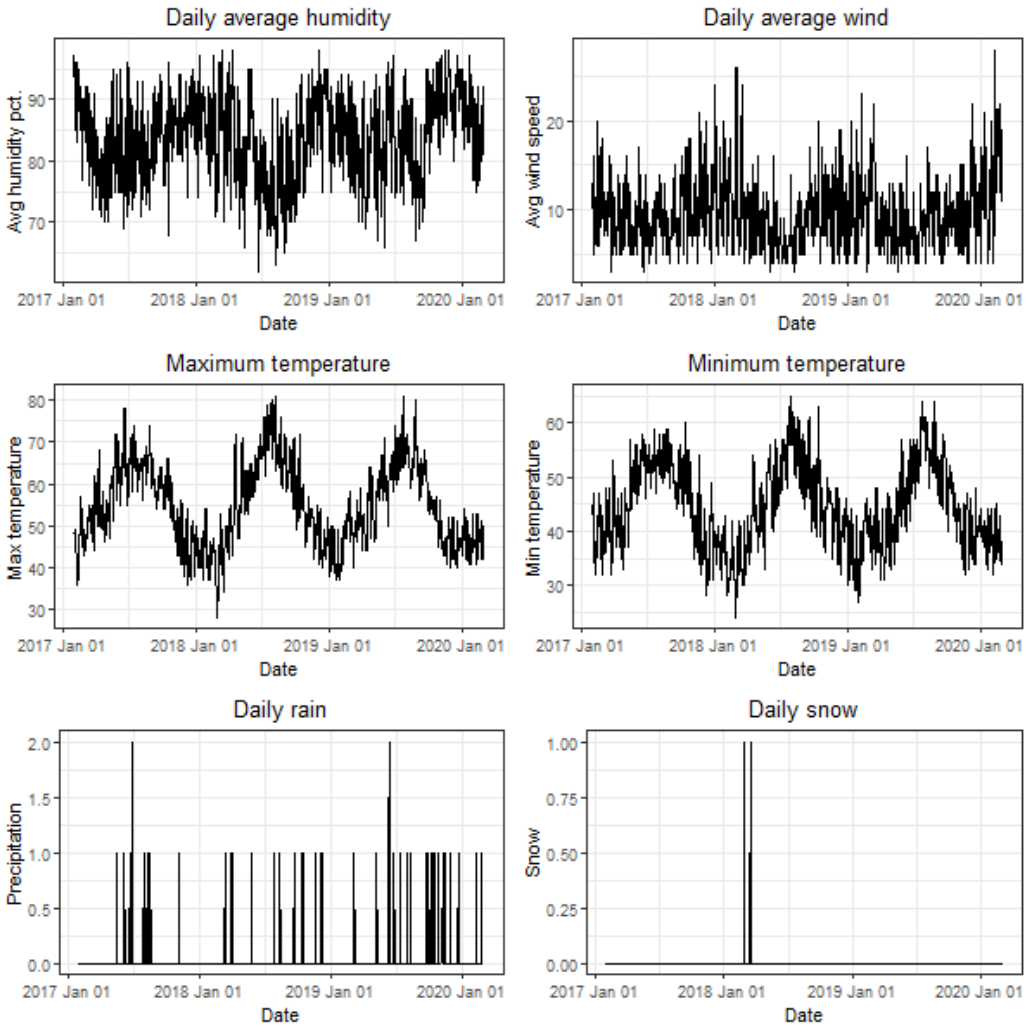

Figure 5: Time series for each weather variable for store 4126.

condition, excluding rain and snow, we thus decomposed the corresponding time series into a seasonal mean, or trend, (computed as a moving average) and deviation from it. The mean was calculated as a 15-day moving average of the corresponding time series. The deviation, or anomaly, was calculated by subtracting the values of the mean from the corresponding original value for each day in the time series. We noted that while sales data only covered the time period February 1st 2017 to 28th February 2020, weather data was provided from February 1st 2017 to July 29th 2021. Figure 6 shows the resulting mean and deviation for humidity and maximum temperature, as an example. Since a 15-day running average does not return a value for the first and last 7 days of a time series, the running average of the evolution of the weather features is only available from February 8th 2017 to July 22nd 2021. Our use of engineered weather feature (see further explanations in the following sections) is thus restricted to match the data for the sales timeseries, i.e. February 8th 2017 to February 28th 2021. Since this data spans just over 3 years, the models discussed later in this report are trained on data that represent all seasons roughly equally.

As mentioned above, we chose a 15-day interval to calculate the moving average for each of the weather features. This choice, although arbitrary, was made while trying to balance two opposing needs: extracting the largest possible deviation while making the trend informative. To illustrate the trade-off between these two, we calculated and compared the average size of the deviation from trend for various window length for the running average. Let say $s(t)$ is the time series for one weather feature at one store and $sma_n(t)$ the moving average of $s(t)$ with a window length of n days, we obtain the deviation $sdev_n(t) = s(t) - sma_n(t)$.

The average size of the deviations for $sma_n(t)$ is then calculated as $m_n i = mean(abs(sma_n(t)))$ for store i. We repeat this process for all 620 stores, obtaining the set of values $S = \{m_n(1), m_n(2) \ldots, m_n(620)\}$. For each n value we compute the mean and the standard deviation. For each weather feature, we repeat this process for $n = 1, 3, \ldots, 59, 61$. This gives us a rough estimate of how informative the deviation from the moving average could be as a function of the window length.

Figure 7 shows that the deviation from trend corresponding to a n-day moving average increases monotonically as n increases, converging to a

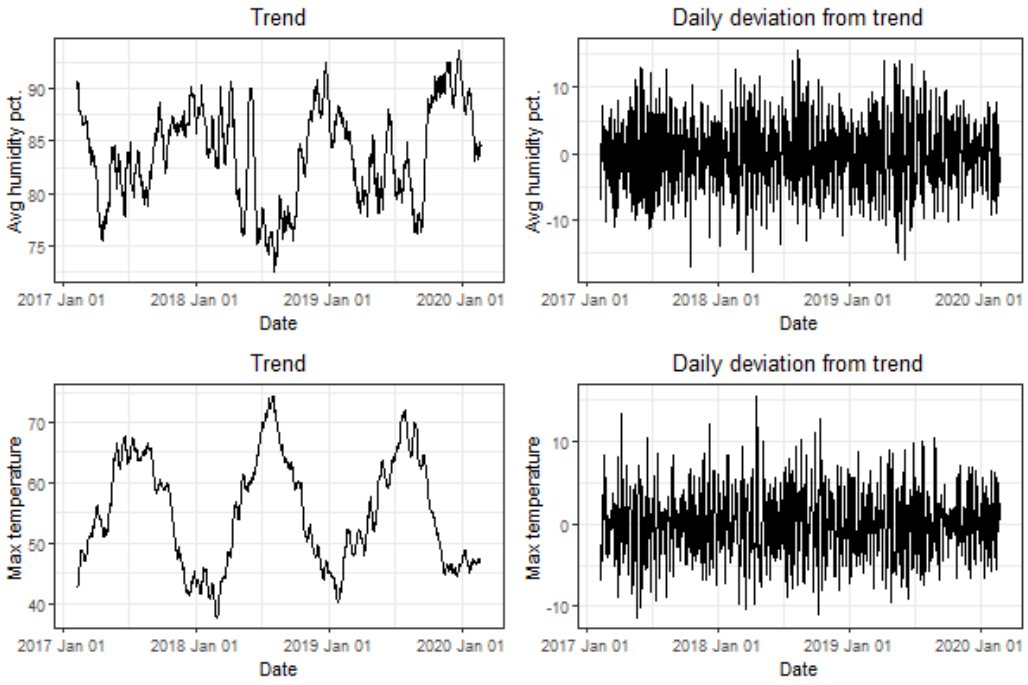

Figure 6: Mean (trend) and anomaly (deviation) of humidity and maximum temperature for store 4126.

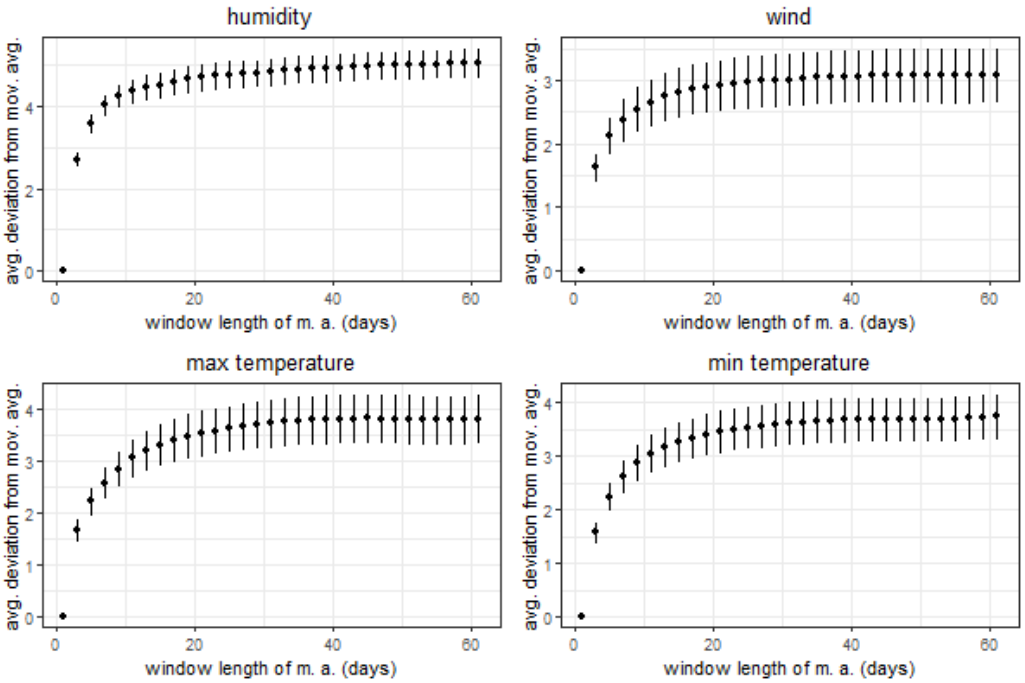

Figure 7: Average deviation from the trend (anomaly) of each weather variable for store 4126 as a function of the number of days n used to compute the moving average for the trend identification.

limiting value as the n-day interval tends to a full year. We decided to use a 15-day moving average as it generally provides a large deviation while we hoped it would preserve local features in the trend. We did not test if our models would yield better results if the engineered weather features were derived using a different window length.

4 Experiments

The dependency of sales on weather is difficult to model for several reasons. In this report, sales of a product is distinguished from demand for that product, with sales being the quantity of goods sold on a particular day and demand being the quantity of goods consumed. Not only are there a large number of covariates which may affect demand, but

there are complicated dependencies linking the covariates, demand and sales. A product sold on day t, is not necessarily consumed on day t, but may instead be consumed on day $t + j$, where j is unobserved. Thus sales on day t depend on the demand for that product on days $t, t + 1, \ldots, t + n$, where n is unknown. For example, if Friday is a cold and rainy day, but the weekend is forecast to be hot and sunny, one would expect sales of certain products such as barbeques, meat and beer to be high. The process is made yet more complicated by the fact that demand on day t not only depends on the weather on that day, but also on the demand on the previous day. There may be a 'first barbeque of the year' effect, where the consumption of a product on a hot day following weeks of cold days may differ significantly from that on a hot day following weeks of hot days. In other words, consumers may get used to the current weather and only change their spending patterns when there is a short-term change in weather.

For these reasons, we decided to tackle the challenge with a variety of approaches having different features. Mainly these can be split into two groups: statistical models and machine learning methods. The statistical models included linear models, generalised linear models and multilevel models. The machine learning approaches included decision trees, random forests and neural networks.

4.1 Linear model

4.1.1 Methodology

As a first step in understanding the relationship between sales and weather, we used a linear model to describe how changes in each weather covariates is related to change in sales for each PPG. Since sales depend on many covariates besides weather, such as the day of the week or the type of store, covariates besides weather were included in the model. Recall the linear model is:

$$Y = X\beta + \epsilon,$$

where Y is a vector of responses, X a matrix of covariates, β a vector of regression coefficients and ϵ a vector of noise, assumed to be normally distributed. We fit independent linear models for log sales of each PPG.

Let $i \in \{1, \ldots, I\}$ index the PPGs so that $Y^{(i)}, X^{(i)}$ and $\beta^{(i)}$ are respectively the response vector, matrix of covariates and vector of regression coefficients for PPG i. Vector $Y^{(i)}$ has a component for each combination of day and store for which sales data is available. That is, element j of $Y^{(i)}$, denoted $Y_j^{(i)}$, is log sales of PPG i in a particular store, on a particular day. The distributions of log-sales appear more Gaussian than the distributions of sales, which may mean the linear model is more suitable for capturing the effect of covariates on log sales. Note that data only exists for day and store combinations which have positive sales, so there is no issue with taking the logarithm of zero. Furthermore it is impossible for the model to predict negative sales since the exponential of a negative value is positive. The downside of using this transformation is that non-integer values are predicted for sales and must be rounded to the nearest integer - for example a log-sales prediction of 1 is equivalent to a sales prediction of approximately 2.7. Let the columns of $X^{(i)}$ be denoted $X_j^{(i)}, j = 0, 1, \ldots, p$, and the elements of $\beta^{(i)}$ be denoted $\beta_j^{(i)}, j = 0, 1, \ldots, p$. The coefficients are as follows:

- $\beta_0^{(i)}$ the intercept,
- $\beta_1^{(i)}$ (categorical) the effect of the store being in Division South, so that the North Division is default.
- $\beta_j^{(i)}, j = 2, 3, 4$ (categorical) the effect of store format, with 'Superstore' taken as default.
- $\beta_j^{(i)}, j = 5, \ldots, 14$ (categorical) the effect of location type, with 'Out of Town' taken as default.
- $\beta_j^{(i)}, j = 15, 16$ (categorical) the effect of Universities, with 'Non University Store' taken as default.
- $\beta_j^{(i)}, j = 17, \ldots, 22$ the effect of day of the week, with Wednesday taken as default.
- $\beta_j^{(i)}, j = 23, \ldots, 33$ the effect of the month of the year, with September taken as default.
- $\beta_j^{(i)}, j = 34, \ldots, 37$ the rolling average of humidity, wind, temperature and precipitation over a 15 day period covering the 7 previous days, the present day and the next 7 days.

- $\beta_j^{(i)}, j = 38, \ldots, 41$ the deviation of the present day weather from the 15 day rolling average for humidity, wind, temperature and precipitation.

The default for all categorical variables apart from Day-of-Week and Month was chosen to be the most frequently occurring category. Wednesday was chosen as the default day of the week because it is not on the weekend and bank holidays are on Monday or Friday. September was chosen as the default month because it is not in the Christmas, Easter or Summer periods. The weather variables were split into rolling average and deviation from rolling average in order to distinguish the effect of short and long-term trends. The coefficients relating to the weather variables are of greatest interest. The columns of $X^{(i)}$ corresponding to categorical variables were encoded as 1 or 0, indicating whether the factor is active or not. The columns of $X^{(i)}$ corresponding to non-categorical variables were standardized, so that for the period considered, each covariate had mean zero and variance one. Importantly, this scaling was common to all models fit, that is, was independent of the length of $Y^{(i)}$. This common scaling helps to make meaningful comparisons between the elements of β. The data was divided into subsets, each corresponding to a particular PPG. These subsets were then divided into training and test sets using an 80:20 split. When a categorical variable was not present in the training subset but was present in the test subset, the data point was simply removed from the test set, so that the linear model was able to make predictions. Removing instances from the test dataset like this is very unlikely to have an effect on the results, because it tends to only be necessary when there are very few instances of a particular categorical variable in the data. The entire procedure takes roughly 40 minutes, a large proportion of which is taken up by creating the relevant subset of data, instead of the actual training and testing phase. The model was implemented in R; the software dealt with rank defficiency automatically, in this case by setting the Regression Coefficient of Location Type 'PFS' as 'NA' (Not Applicable) for all PPGs. Location Type 'Petrol Station' and Format 'Petrol Station' also had 'NA' regression coefficients for a few PPGs.

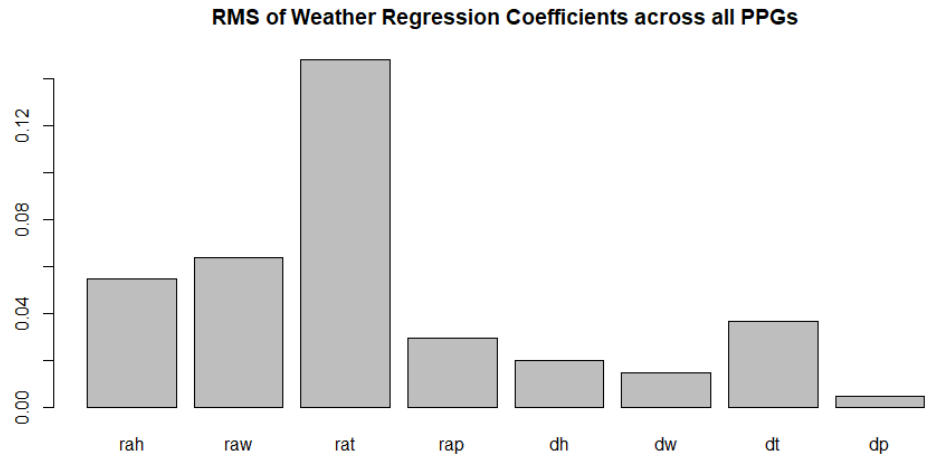

Figure 8: Root Mean Squared of the weather regression coefficients on sales across all PPGs.

4.1.2 Results

To get a sense of the average effect of each weather coefficient on sales across PPGs, we consider the root mean square of each vector $(\beta_j^{(1)}, \beta_j^{(2)}, \ldots, \beta_j^{(151)})$ for $j = 34, \ldots, 41$. The barplot in Figure 8 shows this summary statistic, where labels "rah", "raw", "rat", "rap", "dh", "dw", "dt", "dp" stand for, respectively, "rolling-average humidity", "rolling average wind", "rolling average temperature", "rolling average precipitation", "deviation humidity", "deviation wind", "deviation temperature" and "deviation precipitation". According to the model, the 15 day rolling average has a greater effect on sales than short-term deviations in weather conditions. The model suggests that temperature has the greatest effect on sales.

Figures 9 shows which five PPGs are most and least affected by the weather. The affect of weather is quantified by computing the l^2-norm (also known as the vector-norm) of the weather-related regression coefficients for each PPG: $|(\beta_{34}^{(i)}, \ldots, \beta_{41}^{(i)})|, i = 1, \ldots, 151$. The model finds that some PPGs are far more sensitive to weather than others.

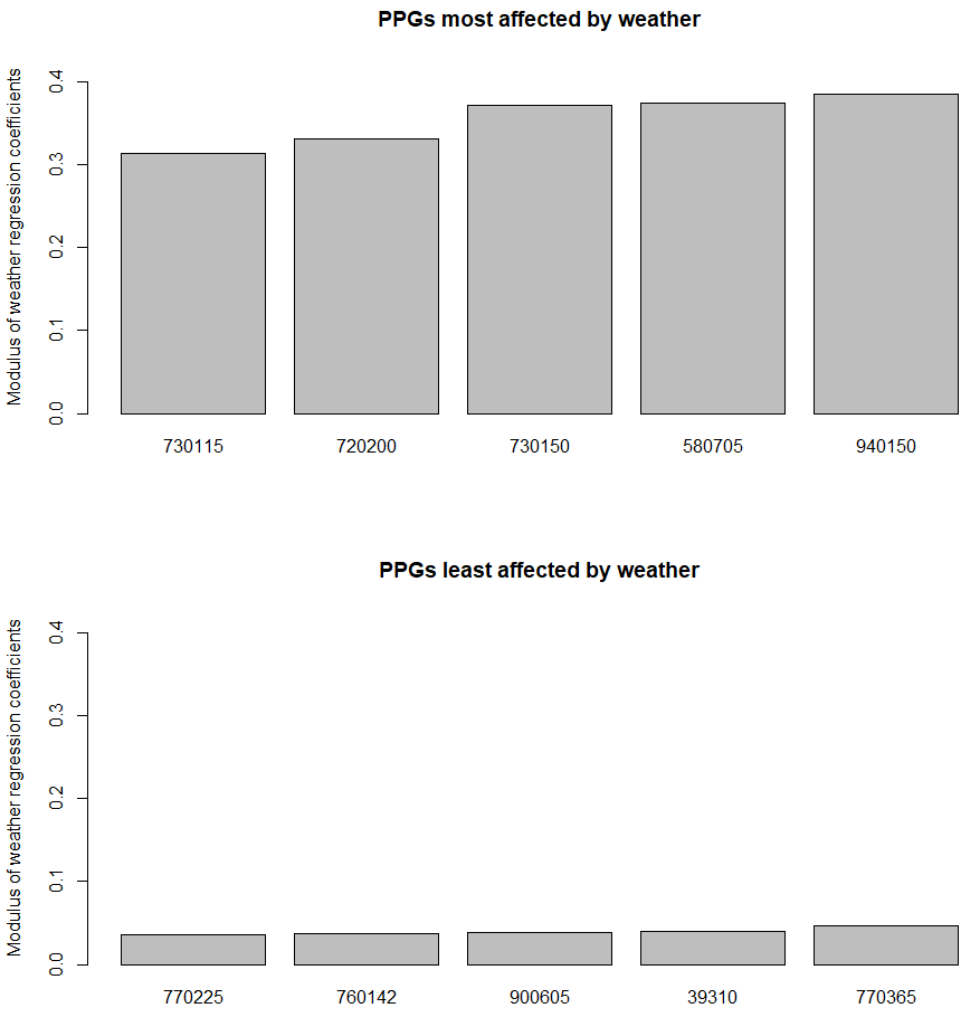

Figure 9: PPGs are most (top) and least (bottom) affected by the weather according to the analysis by means of the linear model.

Figure 10, 11 and 12 show which weather-related regression coefficients had significant t-values, taking as significance threshold 0.001. Black (white) indicates the variable has (not) a significant regression with sales. The plot shows for example that the model finds deviation in precipitation to rarely have a significant effect on sales, but rolling average of wind has a significant effect on sales of most PPGs.

Finally, Figure 13 shows the frequency of root mean squared error (RMSE) of sales prediction for all PPGs, with RMSE not computed on the log scale but on the linear scale. Most of the regressions have errors lower than 30 units. The quality of this prediction should be evaluated with the Challenge Owner on the basis of other pieces of information, such as the storage capacity of each store. Note that this plot doesn't take into account average sales of each PPG. For instance, it could be that some PPGs have a small RMSE only because on average sales are small for that PPG.

The model checks carried out (not shown) show that Gaussian modelling is not consistently a good fit, with some PPGs having non-Gaussian residuals. Moreover, the linear model assumes that training data are independent and identically distributed. This assumption is violated since weather and date are time-series data.

4.1.3 Final remarks

The linear model is simple, computationally cheap, interpretable and well understood, making it a good starting point for analysing the sales-weather relationship. The method used in the initial analysis of the relationship between sales and weather is far from complete however. Simply fitting a single model to each 151 PPGs took about 40 minutes. Improving the model would involve applying a cheap model selection procedure so that the chosen model is better able to generalise to unseen data (thus allowing us for a more robust prediction model).

An alternative approach to making computation on all data more feasible is to cluster the PPGs into 'supergroups', then run a model for each of these supergroups. The plot of pairwise correlations in sales of each PPG (Figure 14) suggests that natural clusters of PPGs may exist. Or course, even if the PPGs could be clustered, each group would contain more data

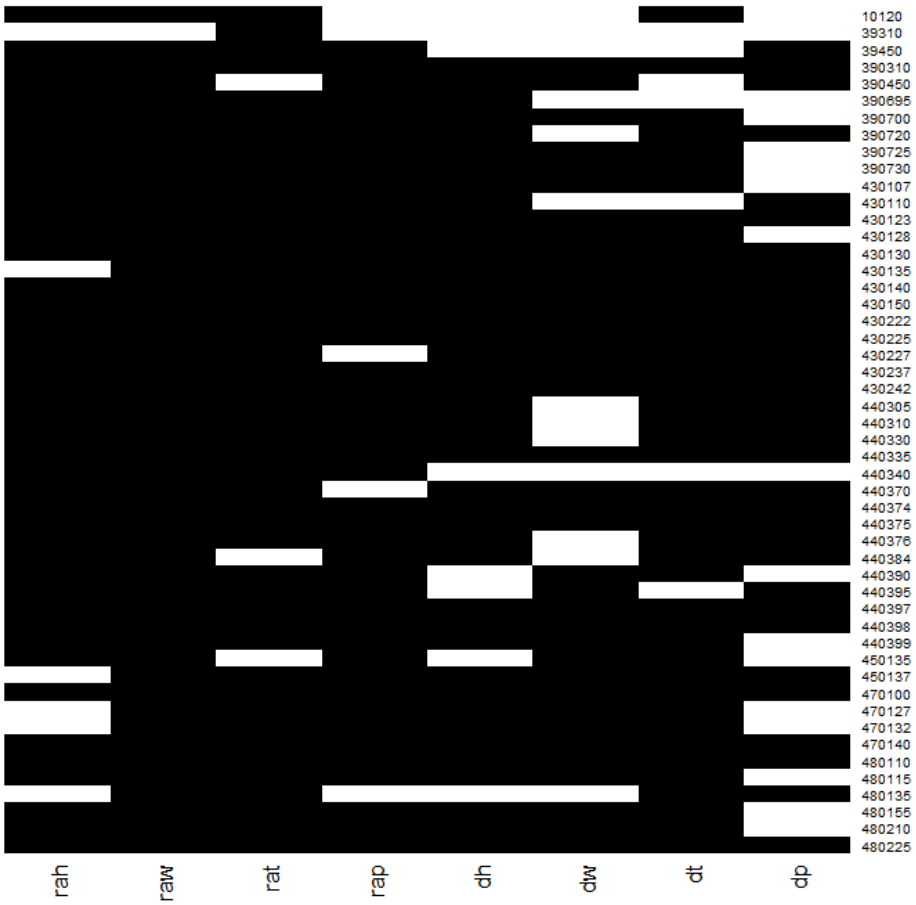

Figure 10: Weather-related regression coefficients having a significant t-values (significance threshold 0.001): black (white) indicates the variable has (not) a significant regression with sales. PPGs are shown on the rows while weather statistics are shown on the columns.

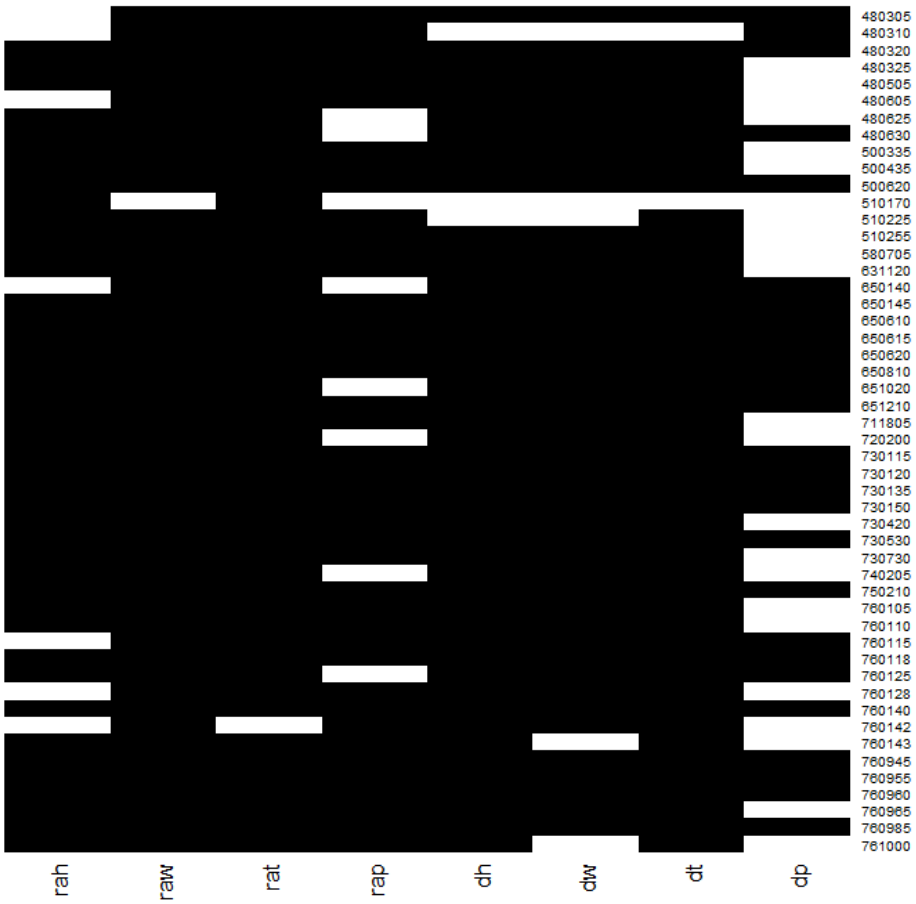

Figure 11: Continued from Figure 10.

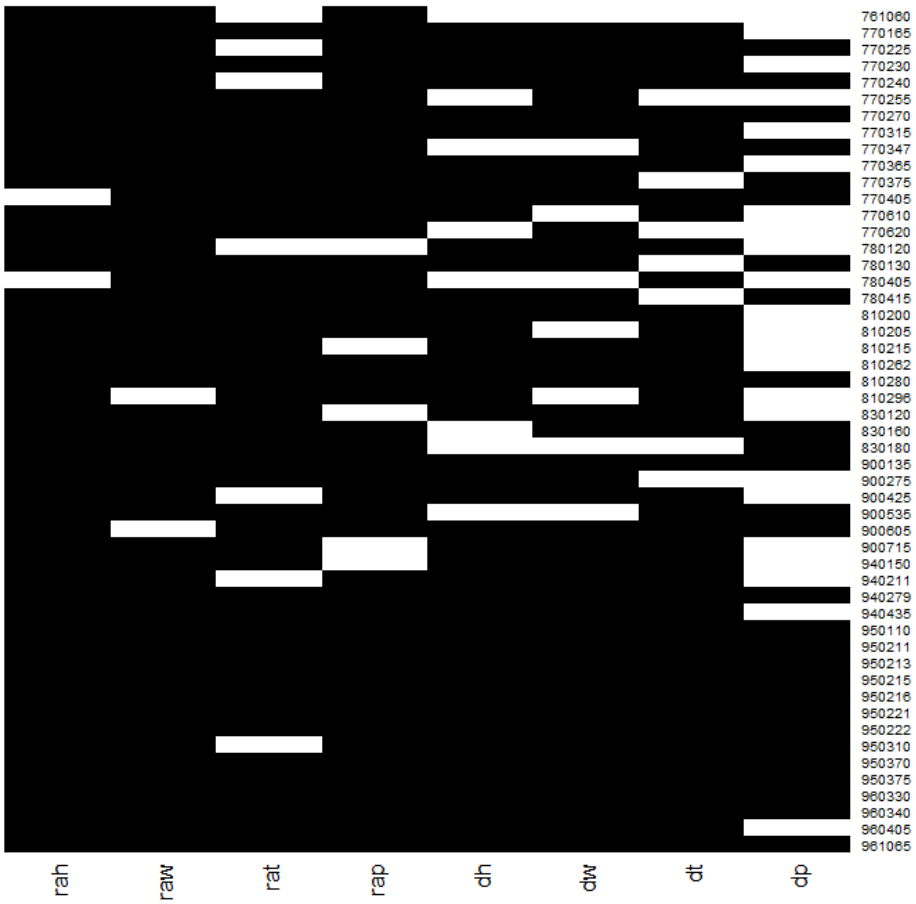

Figure 12: Continued from Figure 10.

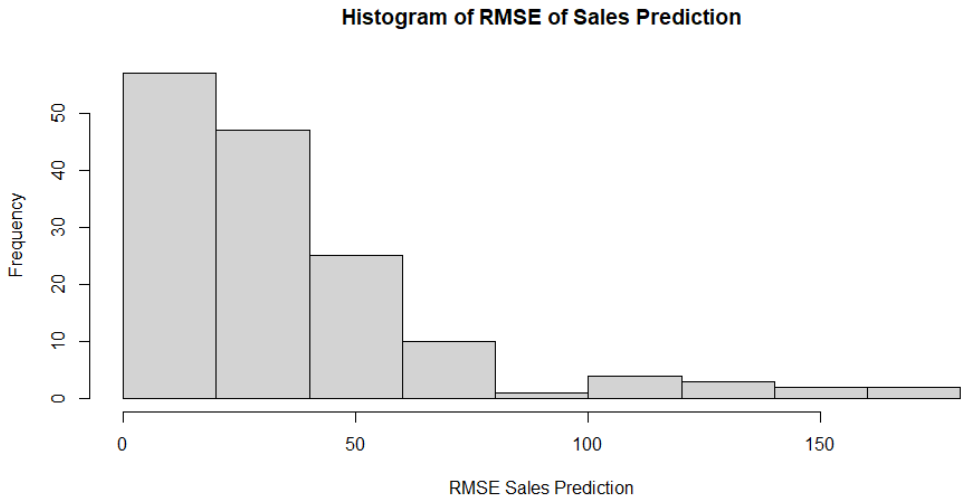

Figure 13: Frequency of root mean squared error (RMSE) of sales prediction using the linear model for all PPGs, with RMSE computed on the linear scale.

points and therefore fitting any model would be more expensive. It is to be tested whether this overall the strategy would reduce computational effort.

4.2 Generalised linear model

4.2.1 Methodology

Generalised linear models are generalisations of the linear model that allow for the response variable to have an error distribution other than the normal distribution. As the sales are discrete values, the Poisson regression or the Negative Binomial models are suitable. The Poisson regression is the best choice if the mean and the variance of the response variable are closer to each other - if they are not and we still use it, this may cause overdispersion in the residuals. The Negative Binomial distribution that does not have this restriction.

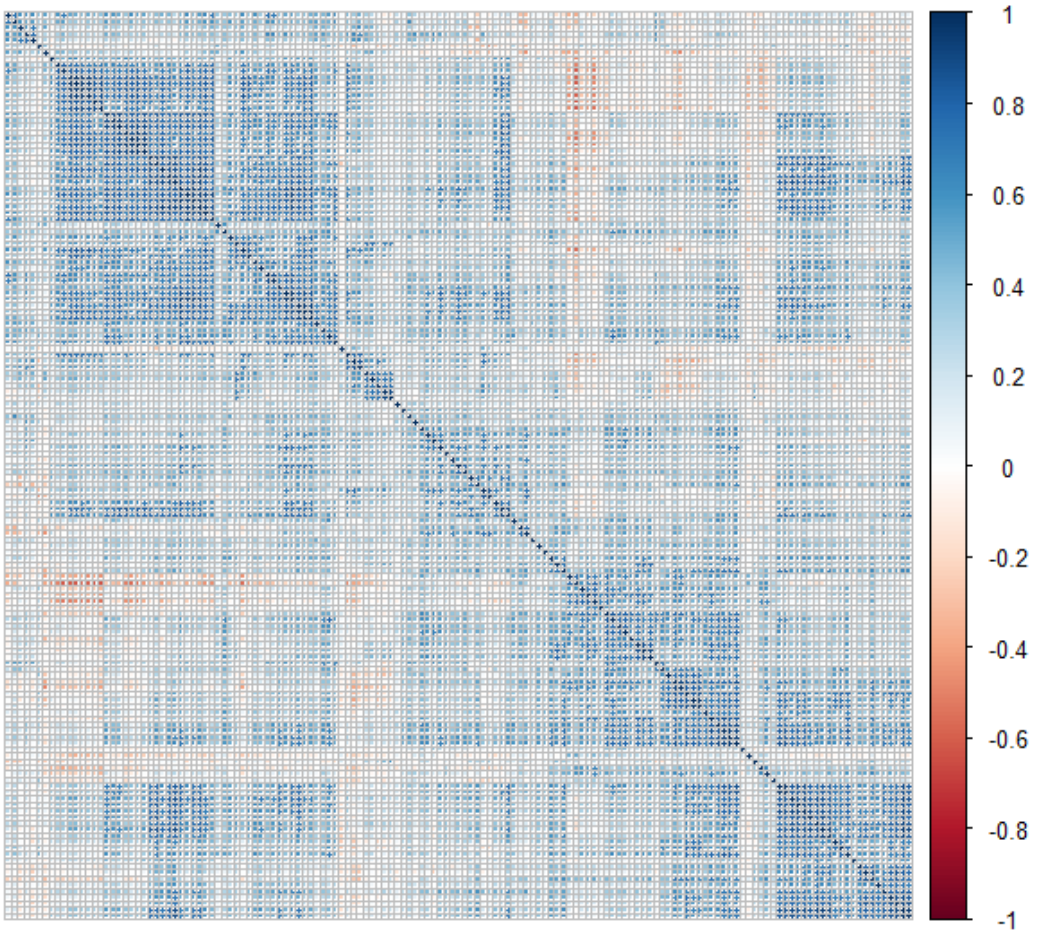

Figure 14: Pairwise correlations in sales of each PPG (note that the plot is symmetrical). Blue clusters may indicate PPGs whose sales time series has similar dynamics and may be analysed together.

4.2.2 Results

For this analysis, we considered the PPG 120 and store 4126. Figure 15 shows the Spearman correlation among all pairs of variables considered (i.e., sales, weather statistics, and time attributes).

Figures 16 compares the empirical frequency of the observed data (in red) and the distributions of a Poisson distribution and a Negative Binomial distribution (in yellow). The Poisson distribution seems a better fit to model the sales data, although this remains to be investigated further and for other PPGs and stores. Table 1 shows the results of fitting a Poisson regression model.

4.2.3 Final remarks

This analysis focused on one PPG and one store only and may provide a narrow outlook on the whole dataset, i.e., the estimates obtained through this analysis may be biased for the population inference.

Nevertheless, this experiment demonstrates that a generalised linear model is an approach that can model the discrete data without resorting to a continuous approximation, as is the case of the linear model. This model can be easily extended to multiple PPGs for one store and also multiple PPGs across multiple stores. In order to account for the heterogeneous behavior of the categorical sales variable and to account for the variation introduced by the weather, seasonality and other covariates, we propose a Multivariate Ordered-response modeling approach next.

4.3 Multilevel model

4.3.1 Methodology

The sales data have a hierarchical, or clustered, structure. For example, we expect week days or weekends to affect store sales in a similar way. Moreover, a multilevel data structure also arises because of the longitudinal characteristics of the dataset, where an individual store's sales over time are correlated with each other. Multilevel models recognise the existence of such data hierarchies by allowing for residual

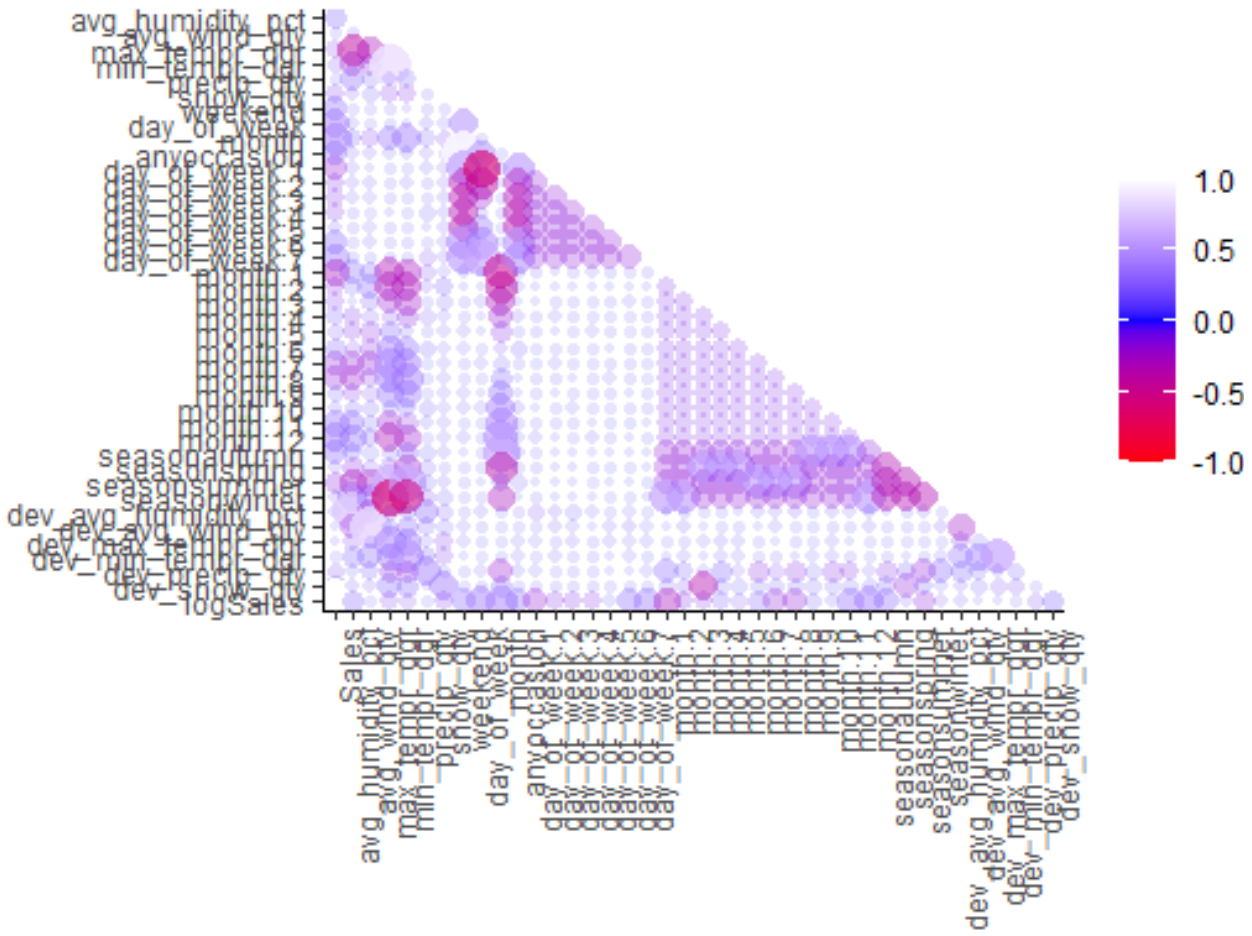

Figure 15: Spearman correlation among the subset of variables used in the generalised linear model. This is a subset of the correlation plot in Figure 14.

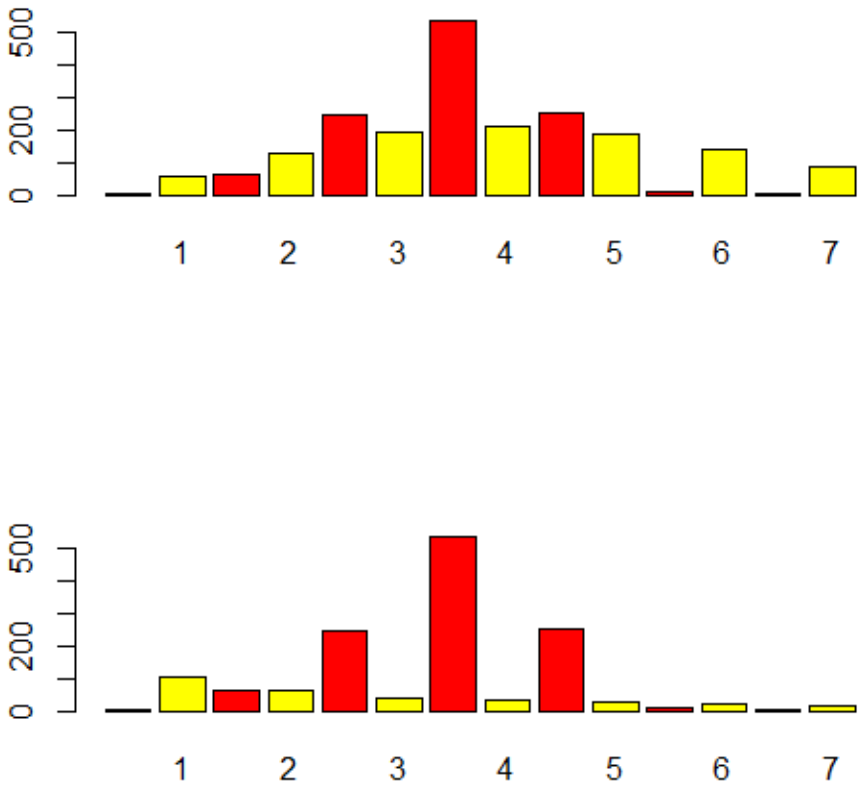

Figure 16: Comparison of the empirical frequency of the observed data (in red) and the distributions of a Poisson (top) and Negative Binomial (bottom) distribution (in yellow).

Table 1: Results of fitting a Poisson regression model. All coefficients are significant but this is expected given the size of the dataset considered.

Predictors	**Incidence Rate Ratios**
(Intercept)	3.59
Average humidity	1.03
Average wind	0.96
Maximum temperature	1.02
Minimum temperature	1.00
Precipitation	0.45
Snow	0.00
Monday	0.85
Tuesday	1.06
Friday	1.19
Saturday	1.70
Sunday	1.75
January	0.48
April	1.15
May	0.80
July	0.51
August	0.53
September	0.86
October	0.91
November	1.42
December	2.89
Deviation from average humidity	0.97
Deviation from average wind	1.04
Deviation from average maximum temperature	0.98
Deviation from average precipitation	2.10
Deviation from average snow	1132.79

components at each level in the hierarchy and thus partitioning the residual variance and obtaining more accurate inference and forecasting.

The levels we considered were PPGs (level one) and stores (level two). In order to be non-trivial there needs to be variation in the levels of sales of PPGs at different stores, which can be seen in the data (for example, 'superstores' have much higher levels of sales of all PPGs on average than 'petrol stations').

The model also allows the regression coefficients to vary across groups in the data. It assumes that coefficients are unobserved random variables drawn from a population. This is useful for the Challenge Owner, because (when ran on a sufficiently sized sample), we would be able to make statements about the population of stores and PPGs without having to include *all* stores and products in the estimation.

The overarching idea of this model is to allow random intercepts to capture underlying levels of sales at the Sales:PPG level (drawing a random PPG from a random store). This would capture unobserved heterogeneity at the store, e.g., consumer preferences or price promotions. Random slopes can capture weather effects specific to each PPG (across all stores). Finally, global slopes can capture control variables specific to all stores and products and unrelated to weather effects.

The framework is appealing as its results are easily interpretable - i.e., an x degree increase in midpoint temperature leads to a roughly $y\%$ increase in sales in product z. Three components are modelled together: the 'global' level effects (e.g. "the day we are forecasting is a Thursday in July"), the performance of the sales of a PPG at each store (e.g. "the store we are forecasting is 4537 and product is 216"), and finally how many degrees from average midpoint weather the forecast is. The results presented in this report are the most computationally efficient: random intercepts are included but random slopes are not, so that the model is computationally tractable.

The covariates included are

- weekday (reference category is Friday);

- average temperature, $Temp$, computed as the daily average between maximum and minimum temperature;
- seasons, s, defined according to the Met office: Spring includes March, April and May; Summer includes June, July and August; and so on.

No other weather variables were included due to time and computational constraints, but the model could be easily extended.

Here we present results from two multilevel models. The first model is defined as

$$\ln(Sales^{p,i,t}) = \alpha_0 + \alpha_1^{p,i} + \beta_1^p \times Temp_{i,t} + \gamma^{i,w} weekday_{i,t} + \epsilon^{p,i,t}$$

where $Sales$ is the daily sales, p is the PPG, i is the store, t is the time, w is the weekday, ϵ is the error term. This model features a global intercept, random effect intercepts for PPGs nested in stores, a global estimate for each of the slopes of calendar controls and slope of weather variables within each PPG.

The second model is equivalent to the first model plus extra control variables related to store type s:

$$\ln(Sales^{p,i,t}) = \alpha_0 + \alpha_1^{p,i} + \beta_1^{p,s} \times Temp_{i,t} + \gamma^{i,w} weekday_{i,t} + e^{p,i,t}$$

4.3.2 Results

Both models are run on a subset of the data that has 15 stores and 15 PPGs chosen at random. The first model has the advantage of being very fast - it takes under 2 minutes to run. Figure 17 shows the global level effects of the calendar control variables. Estimated coefficients in red are negative, blue are positive. All estimated coefficients are significantly different from zero at the 1% level, and confidence intervals are also plotted but are not easy to see. Note that estimated coefficients on weekday are as we have seen in other analyses (the reference category is Friday) with sales highest on Friday and Saturday. Figure 18 show the random marginal effects for PPGs with respect to temperature. One random effect is plotted per PPG (y axis reports PPG number).

The second model is very similar to the first model with some extra global controls relating to store format: DivisionDesc, Format, LocationType,

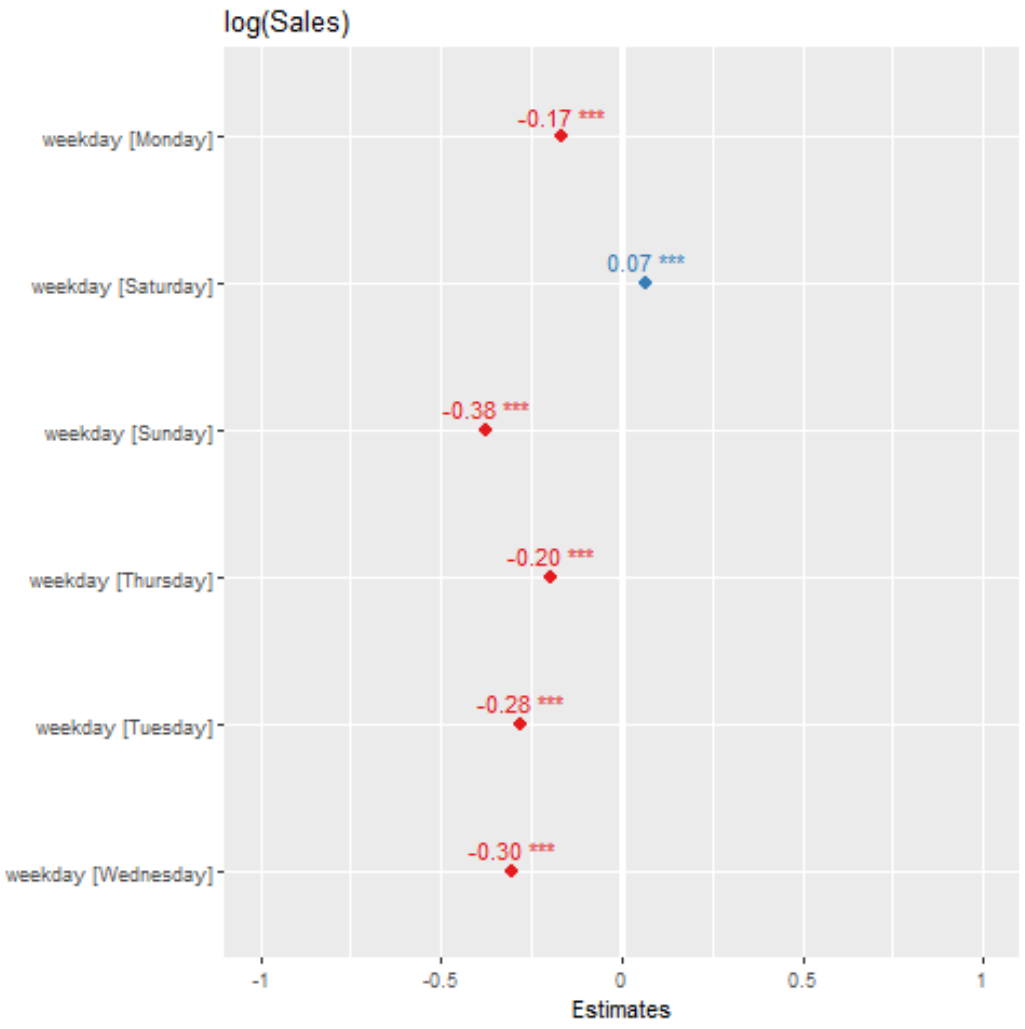

Figure 17: Global level effects of the calendar control variables for the first model.

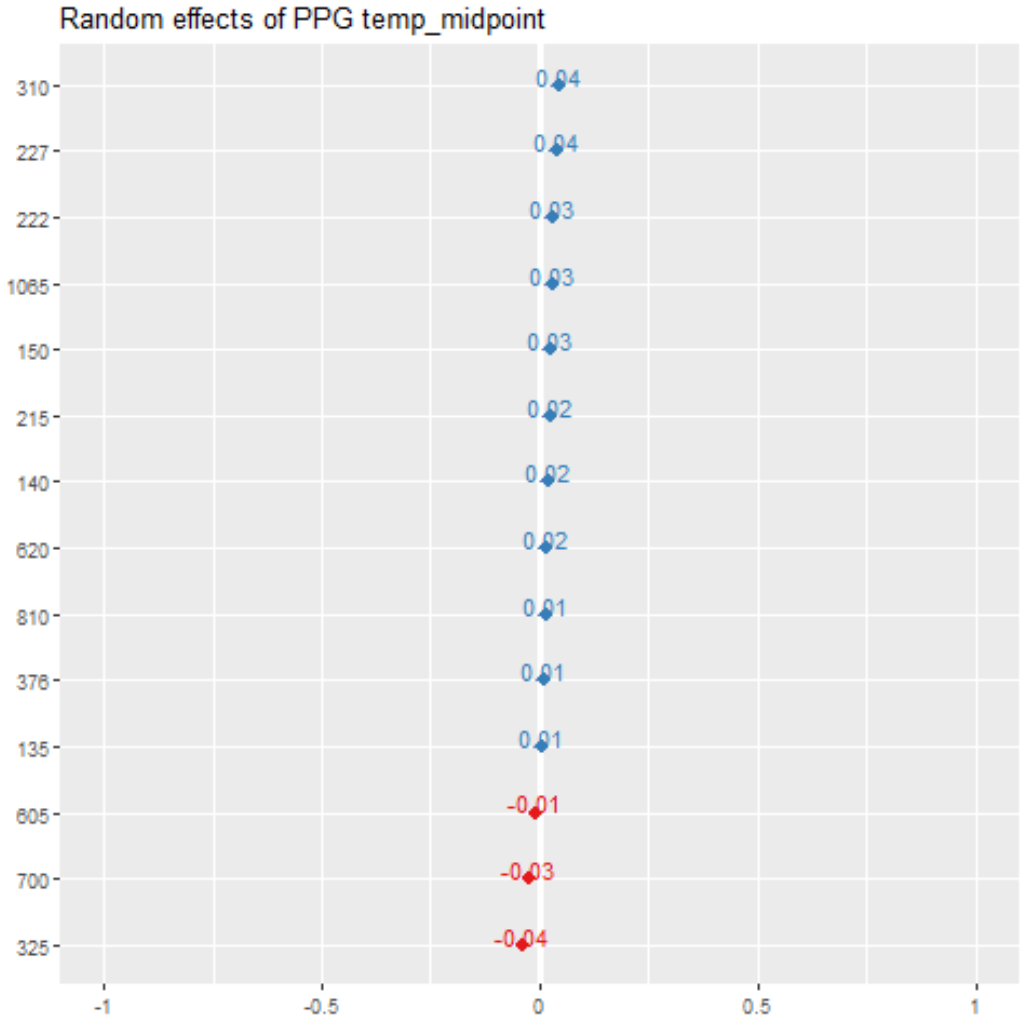

Figure 18: Random marginal effects for PPGs with respect to temperature for the first model. One random effect is plotted per PPG (y axis reports PPG number).

UniversityStore. The model is still very quick to run as it contains no random slopes. The extra controls also add a lot of explanatory power - the R^2 increases substantially. However only the coefficients'confidence intervals for Uni Band B and Edge of Centre, shown in Figure 19, do not contain zero, suggesting they don't add much intuition value for understanding the data.

4.3.3 Final remarks

The major difficulties in implementing this model were the time constraints - to properly specify an appropriate model would take a lot more time and testing than the time allowed in the data study group. In order to circumvent this process, intuition garnered from the other models and data analysis was used to inform the model specification. This experiment showcases to the Challenge Owner the types of features that can be modelled using multilevel modelling and how the results can be interpreted.

4.4 Decision trees and random forests

4.4.1 Methodology

Decision trees belong to the family of non-parametric, supervised statistical methods. They can represent non-linear relationships in the data and can be estimated via efficient algorithms. Decision trees are composed of a set of decision rules which represent the best possible split of the data on the basis of the input variables. One way of representing them is via a set of nodes and edges.

In contrast to classical statistical methods like linear regression, which recover associations between variables, decision trees allow for the prediction of a target variable. They time-efficiently handle large datasets and many potentially correlated input variables. Further, they don't make assumptions about the distribution of the target variable.

The decision trees in this analysis predict the target variable logarithm of sales for a store on a given day. The logarithm of sales was used to account for the non-normal distribution of sales. The algorithm took 62 input variables related to

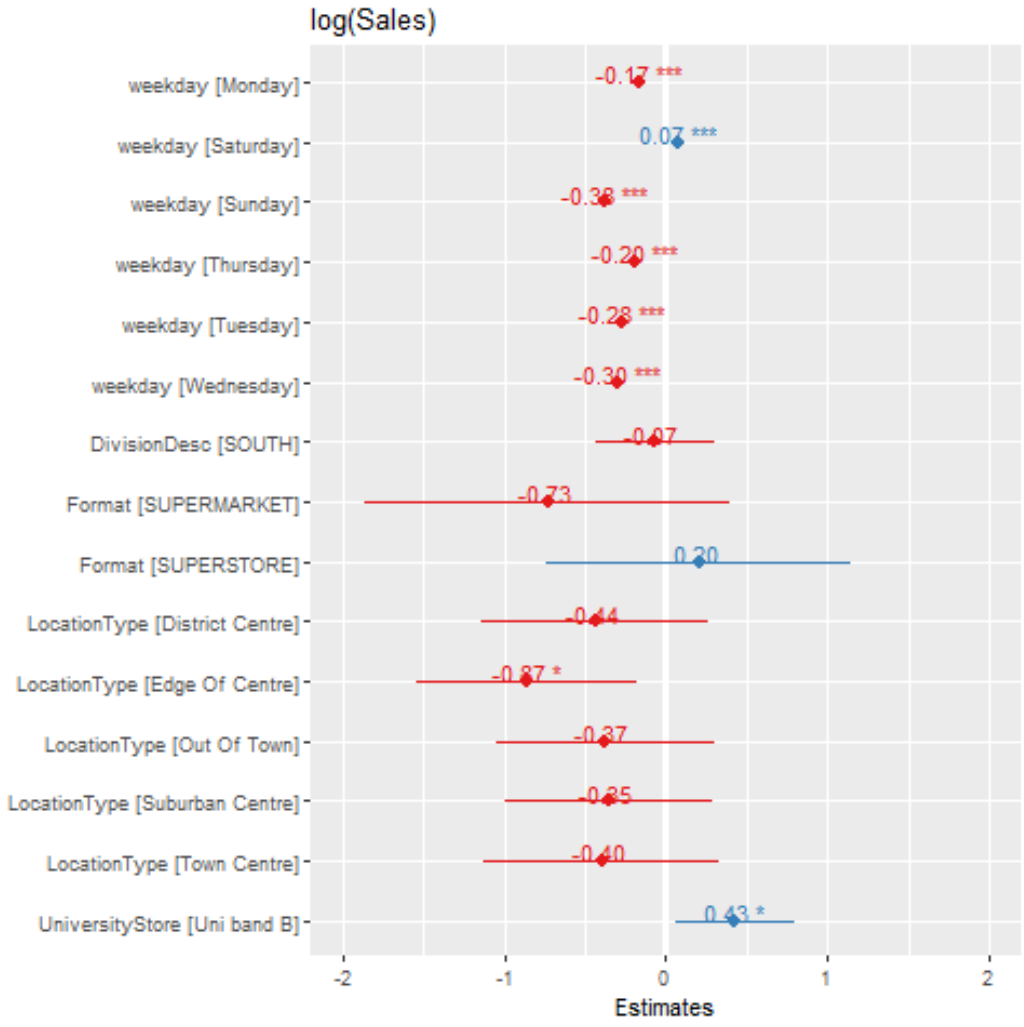

Figure 19: Global level effects of the calendar control variables for the second model.

- weather on a given day: average humidity percentage, average wind quantity, maximum temperature in Fahrenheit, minimum temperature in Fahrenheit, total quantity of rain in inches, total quantity of snow in inches, and the deviations from the moving average of these five weather weather variables
- the store sales for this PPG in this OD were recorded for: division (northern division as the reference category), store format (dummy coded with superstores as the reference category for supercentres, supermarkets and petrol stations), university- or non-university-store (non-university-stores as the reference category), and location type (dummy coded with out of town as the reference category for suburban centre, edge of centre, retail park, district centre, town centre, destination, petrol station, TBC, PFS and null)
- presence of particular events in a given day: whether sales were recorded on a weekend (Friday, Saturday and Sunday) or not, the day of the week they were recorded at (one hot encoded with 1 = Sunday), the season they were recorded in (one hot encoded; March to May = spring, June to August = summer, September to November = autumn, December to February = winter), whether they were recorded on a bank holiday or another day of significance (one hot encoded, included: Boxing Day, Christmas Day, Early May bank holiday, Easter Monday, Good Friday, New Year's Day, Spring bank holiday, Summer bank holiday, World Cup 2018 regular game, World Cup 2018 England Match, none), and whether there was any type of occasion on the day sales were recorded (weekend as defined above, bank holiday or another day of significance)

This simple decision tree has to be read as following: if a node condition applies, go down the right edge coming from this node; if it doesn't apply, go down the left edge. Samples describes the number of samples in that node. Values describe the logarithm of sales predicted for data points for which conditions apply.

4.4.2 Results

The decision tree in this analysis have been 'grown' on data from all stores and one target combination of OD (1) and PPG (120). The decision tree

shows that:

- for big stores, sales on weekdays in December are especially important, not much on weekends or Christmas Day (whereas sales increase a bit on Boxing Day again); other than December, sales are strong in autumn and on Fridays in January
- in supermarkets, it is mostly bought on other days than Sunday and when the temperature drops suddenly in January
- in petrol stations, sales are predicted to be much lower
- sales are marginally greater for the northern divisions
- for the northern divisions, more sales predicted for November when the temperature does not suddenly drop
- for the northern divisions, more sales predicted for December when there is at maximum a light breeze and not too great an increase in humidity which may go together with rain
- for the southern division, petrol stations, no clear predictor of sales increase or decrease relevant for ASDA can be determined

Whereas decision trees are inherently interpretable, their simplicity may come with a loss in accuracy of predictions. Therefore, this decision tree has been trained on all data for the above-mentioned PPG and OD from half of stores and tested on the data from the other half of the stores.

Instead of relying on one single decision tree to output an accurate prediction, we can rely on many of them and average their predictions. This ensemble learning technique is called a random forest. As well as improving accuracy, random forests reduce overfitting in decision trees, they work well in both categorical variable and continuous variable cases, they automate missing value imputation and they are robust to outlier problems in the data.

Due to the high numer of categorical features (i.e., covariates) in our dataset, random forests can help in formulating specific associative rules among the features thus helping us to understand the data.

The table below shows the input variables most important for prediction which can be obtained as a byproduct when training random forests.

Table 2: 20 most important features for predicting log(sales) with a random forest.

Input	Importance
supermarket	0.274
petrol station	0.233
deviation in maximum temperature	0.047
deviation in minimum temperature	0.044
deviation in humidity	0.041
deviation in wind quantity	0.041
January	0.035
maximum temperature	0.034
December	0.029
minimum temperature	0.027
humidity	0.027
wind quantity	0.020
autumn	0.014
Friday	0.014
Saturday	0.013
deviation in precipitation	0.011
university store	0.007
supercentre	0.007
June	0.007
Sunday	0.006

The loss of interpretability when combining the predictions of multiple decision trees comes with an increase in accuracy. This is illustrated in Table 3.

Table 3: Performance of a single decision tree and a random forest in predicting log(sales).

Method	RMSE_log_sales	RMSE_sales	R^2	MAE	Runtime
Decision tree	0.84	83.5	0.54	0.65	1s
Random forest	0.65	59.0	0.72	0.46	6 min 53 s

S.no	% of features considered to construct single tree	No. of trees used	Train Sample Size	Root mean Squared Value	Mean Absolute error	R^2 Score	Max Depth of trees
1	50%	20	538740	0.83	0.70	0.55	10
2	50%	30	606082	0.66	0.52	0.72	15
3	60%	20	538740	0.82	0.69	0.56	10
4	**60%**	**30**	**606082**	**0.64**	**0.51**	**0.74**	**15**
5	70%	20	538740	0.82	0.68	0.57	10
6	**70%**	**30**	**606082**	**0.64**	**0.51**	**0.74**	**15**

1. Saturation of results coming from 4th to 6th observation
2. Highest R^2 score achieved. (The closer it is to one, the better the predicting quality over unknown data)
3. Lowest MAE which is the difference of predicted value from the original one
4. Computationally less costly as compared to 6th observation

Figure 20: Performance metrics showing the regression accuracy for several runs of XG-BOOST.

As well as standard random forest algorithm we also carried out an analysis using XG-BOOST, a state of the art approach for structured data. Written in C++, it is parallelizable and comparatively fast. Measures of accuracy for several runs of XG-BOOST are shown in Figure 20, while Figure 21 shows the most influential features.

Additionally, we used the Shapley Additive explanations (SHAP) as an alternative method to determine the most influential variables. This method connects game theory and machine learning and it is specifically used for interpreting any black-box tree based model. The most important features according to this method are shown in Figure 22. This method is very informative, as it can also identify which variables have positive and negative effects in the forecast.

4.5 Long short-term memory

4.5.1 Methodology

Long Short-Term Memory (LSTM) is a recurrent neural network based architecture which was introduced in 1997 by Hochreiter and

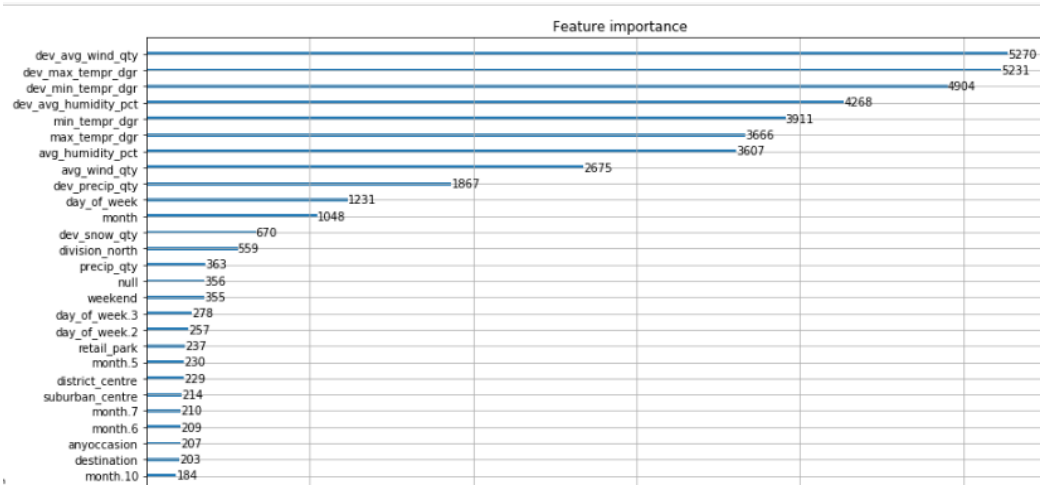

Figure 21: Most influential features as a results of the XG-BOOST runs.

Schmidhuber. LSTM is a high-level recurrent network algorithm whose powerhouse lies in the fact that it possesses the ability to retain information for a longer period of time. This is a major shortcoming especially for older recurrent network algorithms, which has caused them to be slightly ineffective when dealing with large amounts of data with time dependencies.

Considering the huge size of the dataset and also the formidable power of LSTM, applying this technique to learn on the dataset showed an interesting potential. In addition, deep learning techniques such as LSTM have the ability to find interesting features and possibly patterns during training with no need to manually engineer the dataset's features.

Owing to the large data size, which means more data being available to train the model, the model is fed with a lot of training examples which aids the overall learning process. LSTM's memory capacity and ability to store information as time progresses proves useful for this use case. It proves useful because of its nature to learn long term dependencies. The LSTM model stores enough information learned from previous time steps and uses this to inform on future events/predictions.

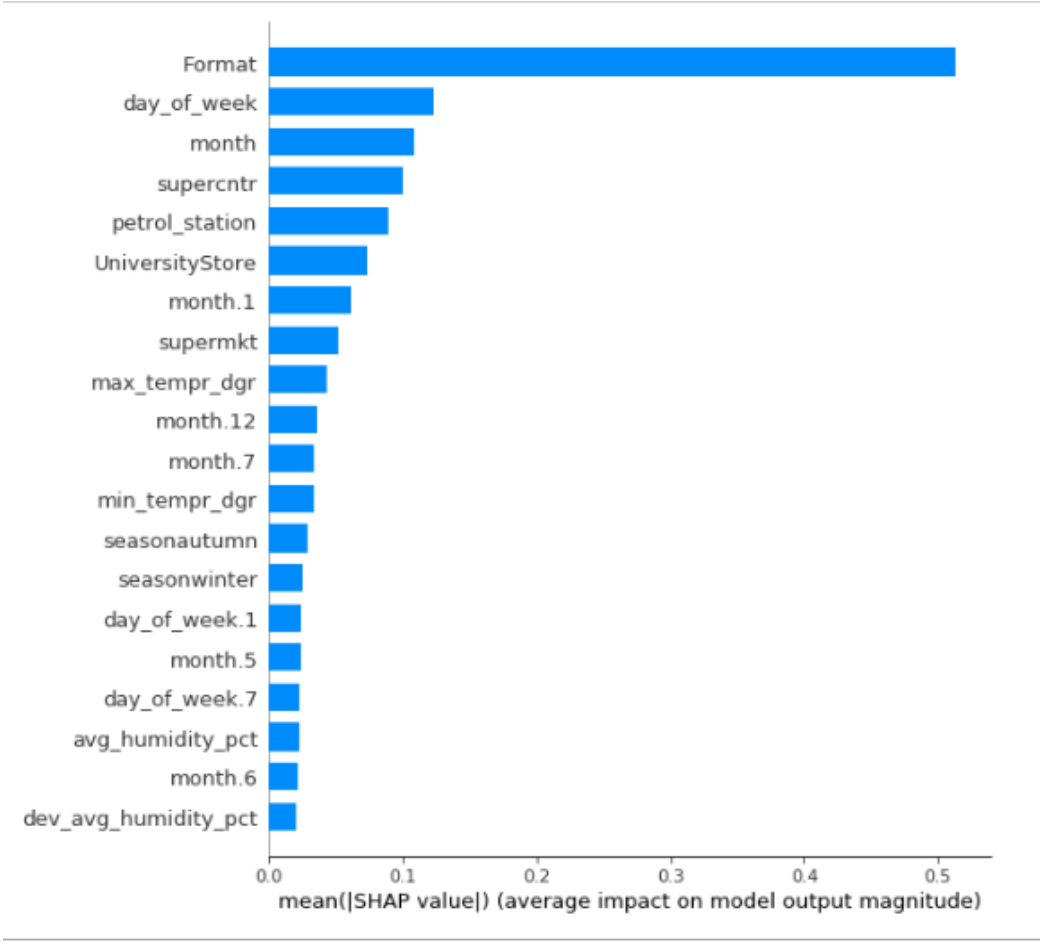

Figure 22: Most influential features according to the SHAP algorithm.

4.5.2 Results

The results shown are for PPG 120 across all stores where that PPG was present. In addition, all categorical features were removed so as to leave the final dataset with only numerical values. The final dataset consisted of 673,425 observations for 53 independent variables and the dependent variable (sales). The dataset was split into training and test, with a 80:20 split, and the data was normalised.

We built the model with 4 LSTM layers/gates having 50 neurons/units each. Also, dropout layers were used in a bid to prevent overfitting. Adding dropouts is a process of dropping random units from the neural network all through the training phase. The dropping of random units during training ensures the units do not acclimate with each other thereby causing overfitting. The model was constructed using a ReLU activation function (activation functions are non-linear properties which are added to neural networks). They help the model learn complex relationships and patterns present in the data.

Two training phases were conducted (this is also known as hyper-parameter tuning). Being a regression problem, Mean Squared Error (MSE) was set as the loss function in both training phases. Several optimizers were utilised to show how tuning parameters can affect the performance of a neural network.

Unfortunately we were not able to achieve model convergence during the short-time of the Data Study Group. However, we believe that this method might be helpful for forecasting.

4.6 Conclusions

We decided to tackle the challenge with a variety of approaches, each having different features. Mainly these can be split into two groups: statistical models and machine learning methods. The statistical models included linear models, generalised linear models and multilevel models. The machine learning approaches included decision trees, random forests and neural networks The limited duration of the DSG challenge (i.e., two weeks) and computational difficulty of the challenge, given by the size of the dataset, did not allow us to analyse the dataset in full, but

rather most analyses focused on a subset of PPGs and/or stores. Different approaches were implemented on different subsets, so this is not a collection of actionable results but rather a review of strengths and weaknesses of each approach which should be further tested on the whole dataset (see more details in Section 5).

Nevertheless, all the analyses agreed that the most influential variables for predicting sales are temperature, day of the week, month, wind and humidity. In particular, it is important to consider anomalies from the long term mean when analysing the weather variables to remove the effect of their natural seasonal fluctuations which may confound the relationship with sales.

The machine learning approaches were extremely fast and useful in gaining an understanding of the relationships between the sales and the other variables but sometimes difficult to interpret. Because the interpretability of the model outputs is very important Challenge Owner, we recommend to adopt approaches which merge the best characteristics that both statistics and machine learning have to offer and provides models which provide an inferential understanding of the relationships in the model, but also improved forecasting ability.

5 Future work and research avenues

There were a large number of ideas generated during the 2-weeks DSG, but only a small number of them were explored. We include here some of the unexplored ideas as well as ideas to improve the methods that we presented in this report.

Clustering the weather variables, the PPGs or the stores might help in identifying additional structure present in the data. For example, clustering techniques could be used to group similar PPGs into one unit so as to create a model dedicated to a specific selection of PPGs. This would also help in understanding PPGs and ODs and uncover associative rules. It may also give a rise to the understanding of weather effects along with the effect of a particular PPG on other PPGs in a specific geographical location. This would give rise to a clearer picture to handle complexities of the unexplained variance when analyzed as a whole.

With regards to the multilevel modelling, the model levels and types of effects can be defined more carefully, considering the input from other models presented in this report. For example, including random slopes on temperature for seasons or non-linear terms. Additionally, multilevel modelling could be combined with generalised linear models, and multivariate ordered response models. This latter model assumes an underlying set of multivariate continuous latent variables whose partitioning maps into the observed set of count outcomes. Based on the observed data, these count outcomes are translated into ordered categories of one, two, three and so-on. Examples of cases where the ordered-response system is used to model count outcomes include household car ownership levels and household purchase of goods, among others. Thus, the joint model system can be estimated using a multi-variate ordered response framework. The ordered-response system allows the use of a general correlation matrix for the underlying latent variables, which translates to a flexible correlation pattern among the observed outcomes. Even though the ordered-response models are used for ordinal responses, the distinction between ordinal and cardinal responses is irrelevant for modelling purposes.

Moreover, we did not consider temporal lags, but we do believe this is likely an important feature for both sales and weather data. This is because, for example, the marginal change in sales due to good weather is likely diminishing - i.e. on day one of a heatwave shoppers will likely want to have a barbecue, but on day five they will be less likely to.

Finally, due to the data avalailable to us, we had to make a range of assumptions. If richer data was available, these assumptions could be relaxed. For example, we asumed that ASDA does not respond to weather in making decisions which affect sales - this likely induces endogeneity bias in the model; understanding the sign and magnitude of this bias and introducing methods to correct for it might be a good avenue for future work.

Furthermore, the response variable (Sales) is truncated, as it is a combined outcome of both demand and stock levels of a product at a particular store. A way to remedy this in future would be to include stock levels in the analysis, or to use a modelling technique that takes into account truncated response variables (such as a Tobit, or Double-Hurdle

model).

Time series of product prices and markups/profit margins would have been very useful in all analyses - for example markups could be used in weighting matrices in order to focus algorithms on the most important parts of state spaces. Prices would also have been useful to transform the dependent variable Sales into a continuous, rather than discrete, response variable. Prices would also be informative of local price promotions. Stock would have been useful to understand demand rather than sales, the latter being truncated whenever demand exceed stock. Finally, richer geospatial information would have been helpful for spatial modeling.

6 Team members

Daniela Anghileri is a Research Fellow within the School of Geography and Environmental Science at the University of Southampton. Her research activity concerns mathematical modelling and optimization for promoting sustainable water management and designing robust climate change adaptation strategies. She was a Principal Investigator (PI) for this challenge.

Giacomo Baldo completed a PhD in the School of Mathematics at the University of Leeds studying evolutionary dynamics and emergent phenomena in collective behaviour. He was one of the participants.

Rachel Joy Forshaw is Assistant Professor of Economics at Heriot-Watt University. She received her PhD in Economics from the University of Edinburgh in 2020. Her current research focuses how macroeconomic conditions affect wealth inequality and labour market outcomes, in particular with regard to the task and skill content of occupations. She examines this field by analysing unconventional datasets using reduced-form econometric modelling techniques. Rachel facilitated this study group.

Franziska Günther is a PhD candidate with the University of Manchester and the Breaking Free Group. Her work focuses on computational models of (self-reported) user data accruing from the use of a digital therapy app

targeting substance abuse. She has read Psychology at the University of Konstanz and Amsterdam. She was one of the participants.

Israel Ilori is the Lead Machine Learning Engineer at TruTalent. He also consults and leads on Data Analytic projects at Tranzfar Limited (a money transfer service based in London, UK). He was a pioneer on the Masters in Data Science program at the University of Manchester. He was one of the participants.

Guneet Singh Kohli is a final year undergraduate student at Thapar University, Patiala studying Computer Engineering. He is an undergraduate researcher with experience in various domains of Machine learning and Artificial Intelligence like Computer Vision, Reinforcement Learning, Natural Language Processing, Machine Learning Modelling, and Analysis. He is currently working on various research projects and looks forward to making contributions to the community with exceptional results. He was a participant in the DSG.

Silvia Liverani is the Head of the Statistics and Data Science research group, Reader in Statistics at Queen Mary University of London and Turing Fellow. Her research interests are in Bayesian statistical models and her expertise is on clustering methods and spatial modelling. She was a Principal Investigator (PI) for this challenge.

Hector McKimm is a PhD student at the University of Warwick supervised by Professor Gareth Roberts and Dr Murray Pollock. He is on the Oxford-Warwick Statistics Programme, a Centre for Doctoral Training run by the Universities of Oxford and Warwick. His research interests are in Computational Statistics. He was a participant in the DSG.

Abhilash C. Singh is a Transportation Engineering Ph.D. student at Urban Systems Lab, Centre for Transport Studies at Imperial College London. His research aims to solve for endogeneity and choice set issues, especially their applications to residential location choice models with a large number of choice alternatives. Concurrently, he is also interested in policy-based analysis to examine equity of access considering the safety and quality of activities accessed by developing novel accessibility measures for Pathways to Equitable Healthy Cities project. In the past, he has worked extensively in understanding human travel-behaviour through development and application of econometric and

statistical methods. He received a Master of Science in Engineering with a specialization in Transportation Engineering (2017) at The University of Texas at Austin, USA, and a Bachelor of Technology in Civil Engineering (2016) from the Indian Institute of Technology Bombay, India.

turing.ac.uk
@turinginst