1. The active members of your group and a brief summary of their contribution to the project and writing of the report.

Isaac - Exploratory Data Analysis, Report Writing, Model Creation and Analysis.
Josh - Exploratory data analysis, Report writing, Model Creation and analysis, collating all code.
Oscar - Exploratory Data Analysis & Report writing & future data suggestions
Louise - Exploratory Data Analysis, report writing, compiling report in presentable format, assisting with various modelling techniques used and future data suggestions

2. A brief description of your project and details of the supplied data. (1-2 pages)

Rossmann Pharmaceuticals, commonly referred to as Rossmann, is a large chain of drug store in Europe with over 4000 stores spanning 8 countries. In 2015, Rossmann enlisted the help of freelance data scientists on Kaggle to assist in the prediction of store sales, providing detailed, daily information about 1115 stores.

The Forecasting sales project had the goal of trying to accurately predict sales for Rossmann stores up to 6 weeks in advance. 3 different csv files were supplied, each providing different information. A ‘train’ csv that contained the daily information regarding each store was the main file aimed to be worked with. This csv was about 1 million entries in length, providing a significant backing for analysis. The information provided detailed things such as customer counts, sales figures, if the store was open, ongoing promotions as well as other basic store information.

Another csv, ‘stores’ provided a greater amount of detail regarding each unique store. This csv was 1115 entries long – one for each store. The details included were mainly information about the nearest competitors as well as other geographical features. Another csv file, testing.csv, was provided in the contest and was what Kaggle used to determine the winner of the contest. The testing csv file contained all the same data points except sales were removed. Due to this, sales were attempted to be predicted with customer numbers, in line with the contest. Even if the real-life application wouldn’t contain customer numbers.

3. Details of any preprocessing that was required prior to loading the data into Python. (1-2 pages)

Little pre-processing was required prior to loading the data into Python aside from understanding the dataset, the data was provided in two separate datasets ‘train’ and ‘store’, a ‘test’ dataset was also available however was not used for the task at hand. Both datasets used were in formats that could easily be interpreted in python, thus, there was no need for any editing or slicing.

4. Details of the processing and manipulation of data in Python. (1-2 pages)

Initial viewing of the dataset revealed a few issues that needed to be addressed. The first step taken was to perform an inner merge on the two relevant datasets, providing a complete collection of daily sales figures matched with store information.

Upon analysing the data types of each feature, it was found that a binary column for state holiday contained many string 0's instead of integers. These were replaced with the proper data type, allowing for proper analysis later. While the 'train' dataset appeared complete and contained 0 NaN values, the 'store' dataset contained missing values in 6 columns. It was decided that only 3 of the features could be salvaged, due to a significant proportion of data missing in the others. The 3 remaining features with NaN entries would have values imputed through kNN imputation. As kNN algorithms struggle with high dimension data, only features with a noticeable correlation were selected to help input the data.

While the 'train' dataset initially appeared complete due to the lack of NaN values, it was found when grouping stores by their store number that approximately 13% of stores were missing data over a certain period of time. After conducting further investigation, it was found that certain stores did not collect data over a 6 month period in late 2014. It was decided that only stores that contained full, complete data would be analysed moving forward.

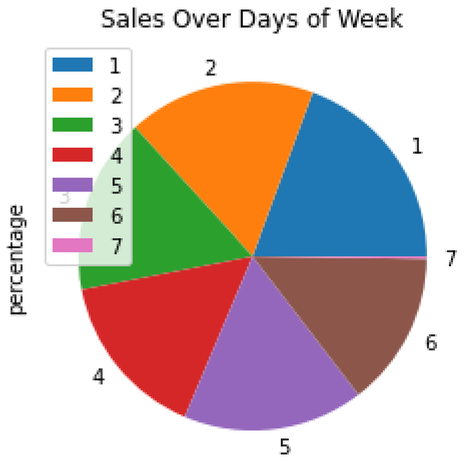
The train data set was 1017210 entries long and thus would be far too computationally intensive to work with. To best manage this the data was sorted by sales and the middle 20% of data was used for future analysis, 10-30% and 70-90% of data was also investigated to determine the best split of data to use however the middle 20% provided the best spread of data without losing a large portion of it. Sorting the data by day of the week revealed very little sales on sundays, this was due to most stores being shut on sundays so in order to construct a proper analysis all of the sunday data was removed, another issue faced is a large amount of missing data in some of the stores in the middle of 2014, to avoid this only stores without missing data were used. The final step of processing and manipulation of data was merging the store dataset with the train dataset.

Initially, the dataset was indexed by store number. For our purposes, it would be more advantageous to have store numbers as a feature. The date column was changed from string data type to a timeseries. Allowing for it to be used as a useful index. Furthermore, it allows for the creation of new columns 'month' and 'day of month'. These features will allow for an analysis of sales over certain periods of time, increasing the performance of our models. The data also contained two columns with categorical variables that required one hot encoding before any analysis could be performed.

As the aim of the project is to predict sales up to six weeks in advance, the data was split into two data sets for testing and training of the algorithm. The testing set compromised the most recent six weeks of data, while the training set contains the rest of the data. As the dataset spans a period of 3 years, this proportion leads to a train / test split of 0.9425. It should be noted that while only 20% of the original dataset is being analysed, there are still over 200,000 entries. Many machine learning models require significant computing power, especially when iterating over high dimensional data. Due to this, for some models required high levels of sampling prior to analysis.

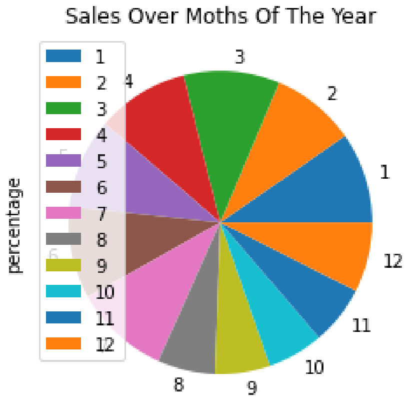
5. A summary of the exploratory data analysis that was performed with any significant conclusions that were obtained from this analysis. (1-2 pages)

When conducting exploratory data analysis, a particularly important aspect of focus was the effect of time on sales. Generally, sales follow yearly cyclical trends that reflect significant moments of modern society. To fully understand the effect of such trends, plots were made to see the effect of sales over different periods of time.
First, it is important to understand how sales differ on a day to day basis. When graphing sales over different days of the week, certain conclusions can be drawn.



Most notably, sales do not differ significantly from day to day. Over the course of the week, total sales do not differ by more than 2% going from one day to the next. There is one notable exception however. It can be seen that Sundays provide an incredibly small proportion of total sales. This can be attributed to the vast majority of Rossmann stores being closed on sundays. Due to this fact, all Sundays will be assumed to have 0 sales, as that is representative of the large majority of data.

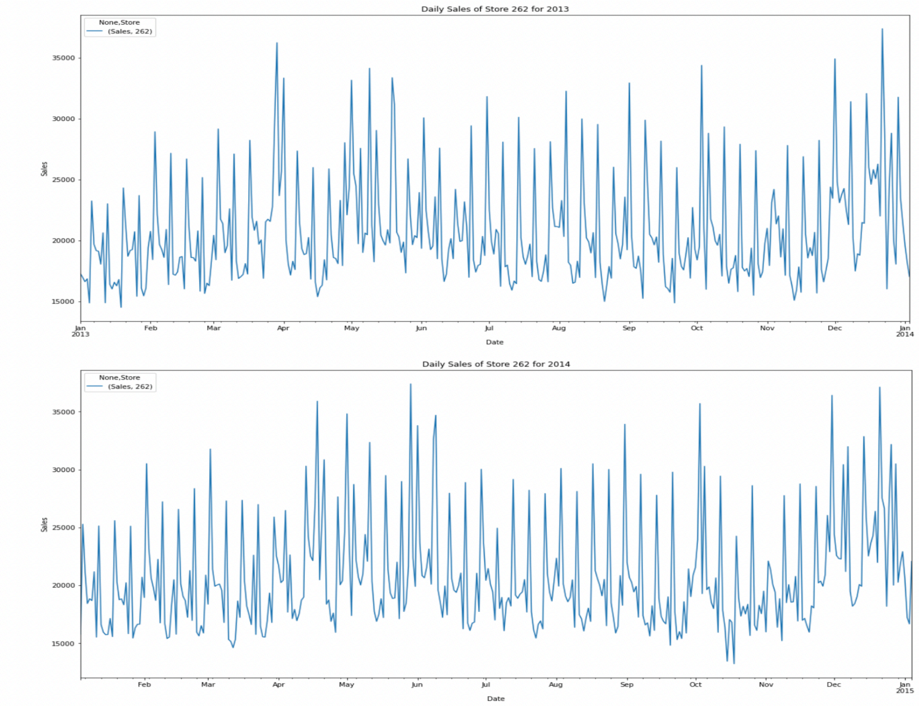
Next, sales were plotted over the months of the year:



Similar conclusions to the days of the week can be drawn from this graph. However, there is a slightly larger variation in the sales figures over the months, with August through February having slightly larger sales figures compared to the rest of the year. This is in line with expectations as generally, pharmacies are busier in winter months. Furthermore, it is recognised that sales are boosted during the holiday period. This is also expected as larger cash flow is more likely during this time.

Simple time series plots and rolling mean plots were investigated to get initial ideas into the effect of the day of the week and time of year affecting sales. To get smooth graphs we had to remove any sales values of 0, indicating when the store was closed. The first store explored was store 262, the top performing Rossman store. Figure 1 shows two years worth of daily sales. From this figure an initial idea into the effect of the december and january period on sales and how the day of the week may impact the stores was found

Figure 2 shows another two years worth of daily sales, this time for store 1114, which was chosen at random as an above average performing store. The sales appeared to be more consistent and less volatile compared to the top performing store. Similarly the sales throughout the 2 years almost mirror each other exactly. This steadiness throughout two different years could be an indicator that the store doesn't seem to have any apparent growth over the years so the year of the sales can be ignored as a variable. To further look into this, a 15 day centered rolling mean for the stores was also investigated.

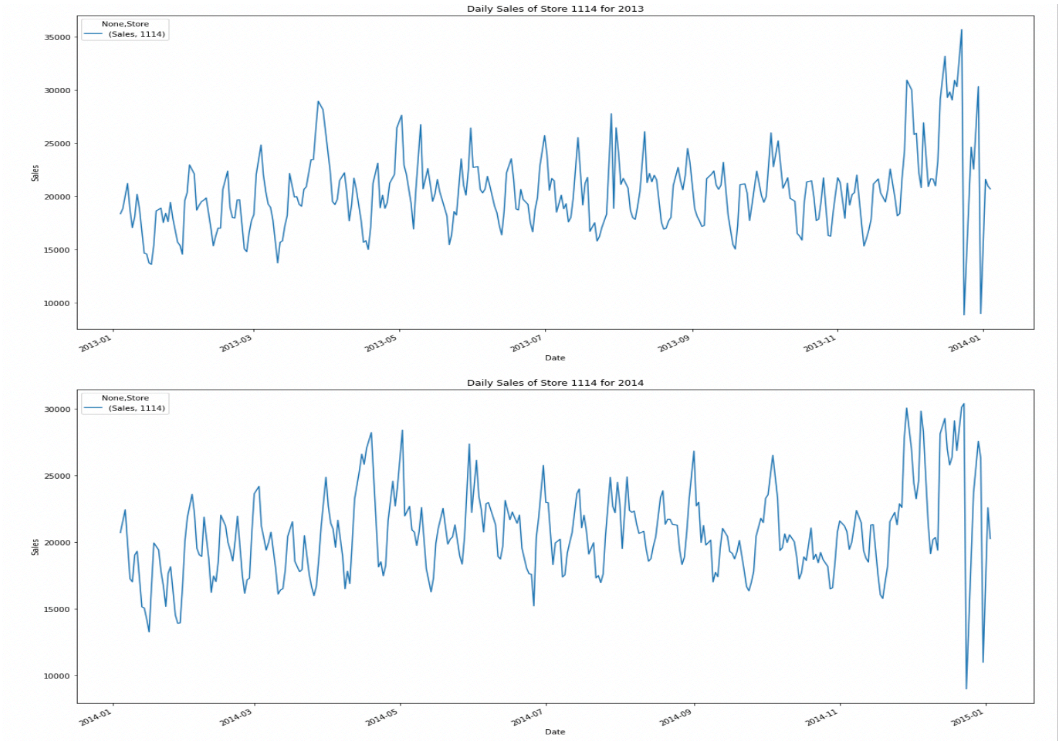


Daily Sales of Store 262 for 2013

Daily Sales of Store 262 for 2014

Figure 1:

Figure 2:



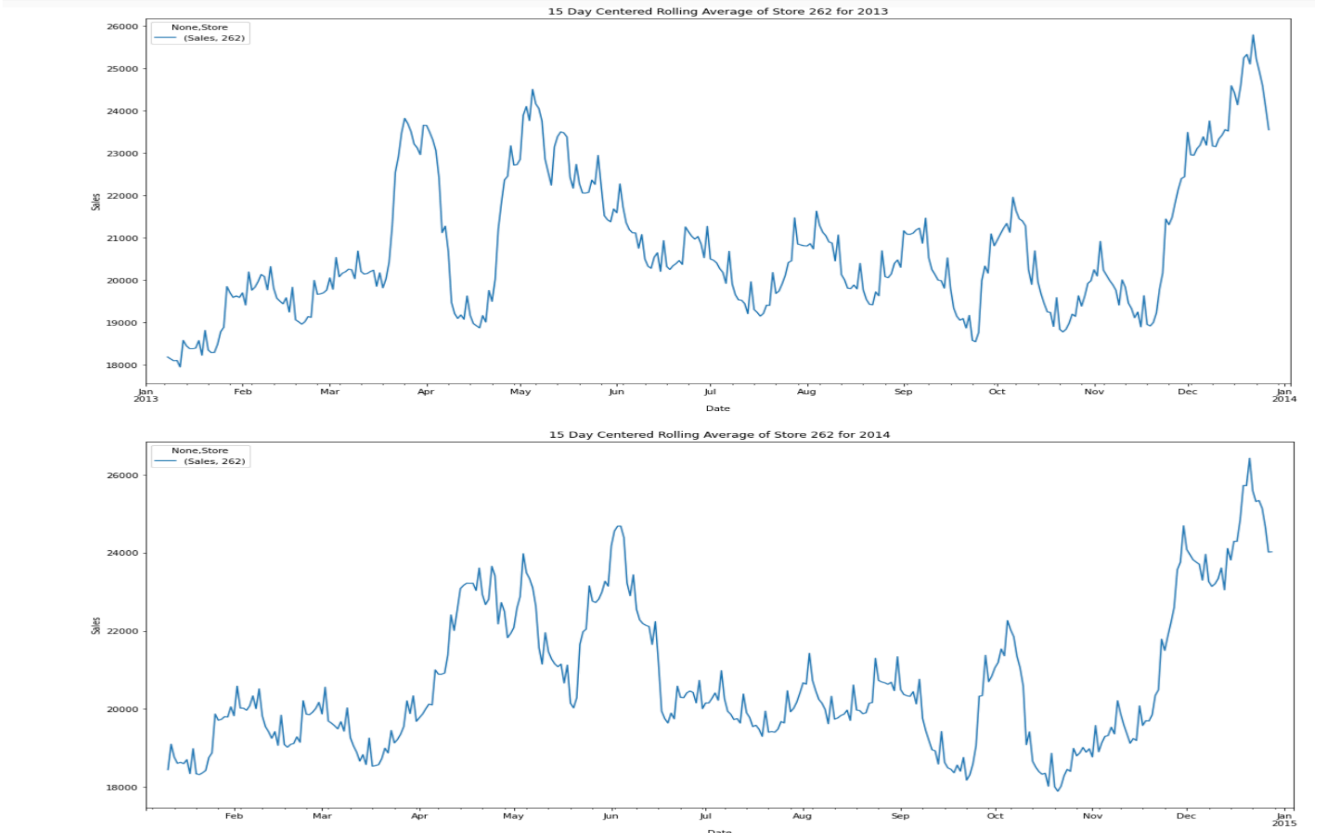
Daily Sales of Store 1114 for 2013

Daily Sales of Store 1114 for 2014

Figure 3 contains the fifteen day centered rolling mean sales for store 262. This was explored to potentially address any major trends that may arise and factor that into later modeling and analysis. The fifteen day mean was chosen as a number that would represent a three week period with the store closures and removal of these days. A trend arises throughout both of the years as there is a consistent rise around December and into January. This can be accredited to the holiday period so this increase may mean that time of year is an important factor when trying to forecast future data. The importance of months was contradictory to the previously mentioned data analysis on months. This is because the previous analysis only took total sales for the months and factored in the days the stores were closed into the average. From this paired with the time series analysis. We can deduce that stores are open less throughout december but when they are open, they tend to have higher sales

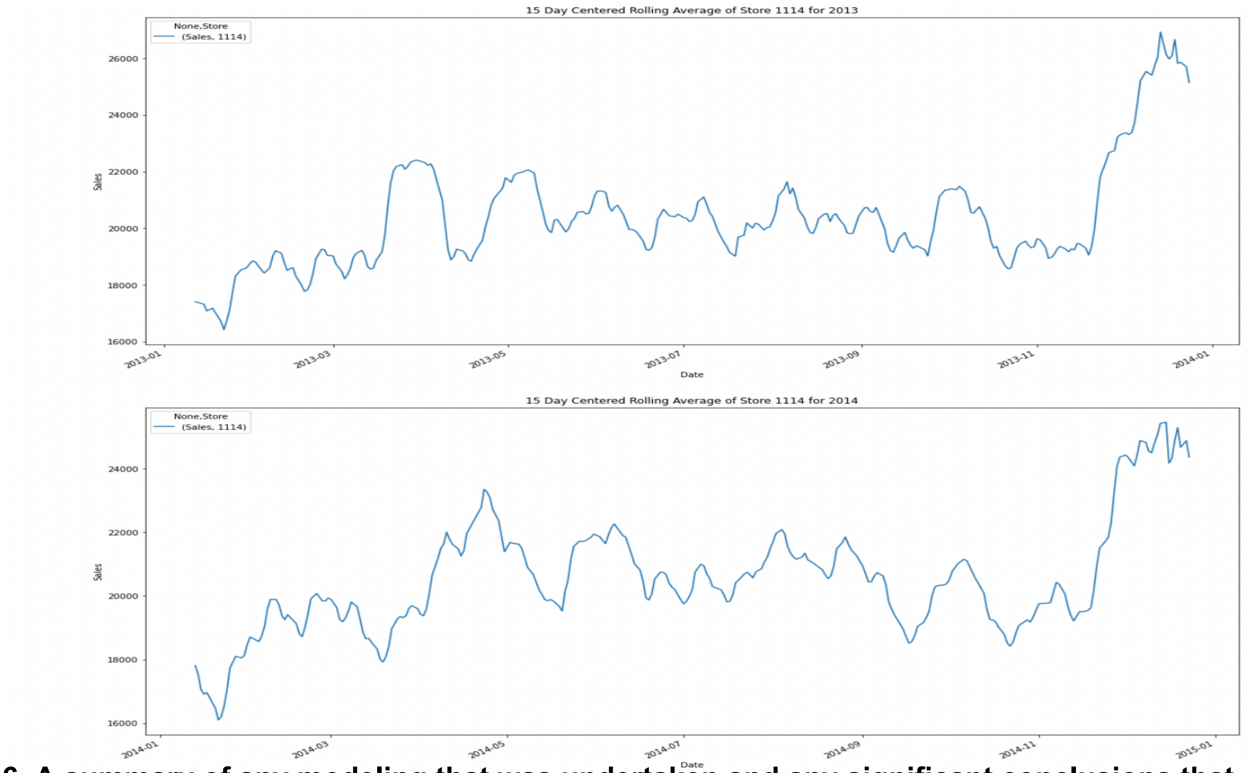
Figure 4 again shows the fifteen day centered rolling mean for store 1114. The two years are very similar with again a rise around the holiday season of December and into the beginning of January. A conclusion that was drawn was that there seems to be no apparent growth across any six week period across the year, apart from the December period. Also there is no growth between the two years with them being very similar throughout. This means that when models were trained, yearly growth was not considered to be something that could have an effect on sales.

Figure 3:


15 Day Centered Rolling Average of Store 262 for 2013

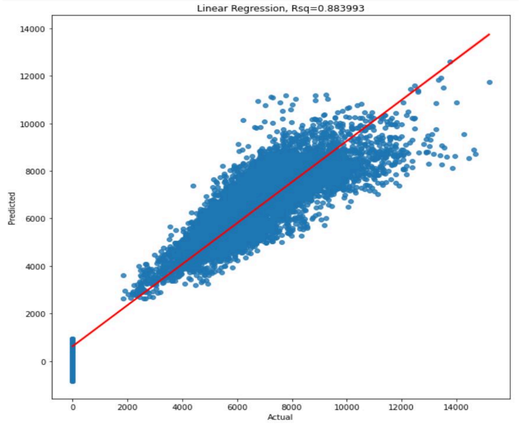
15 Day Centered Rolling Average of Store 262 for 2014

Figure 4:

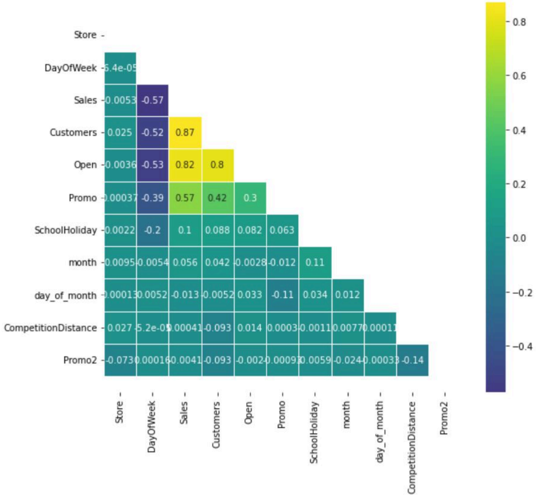


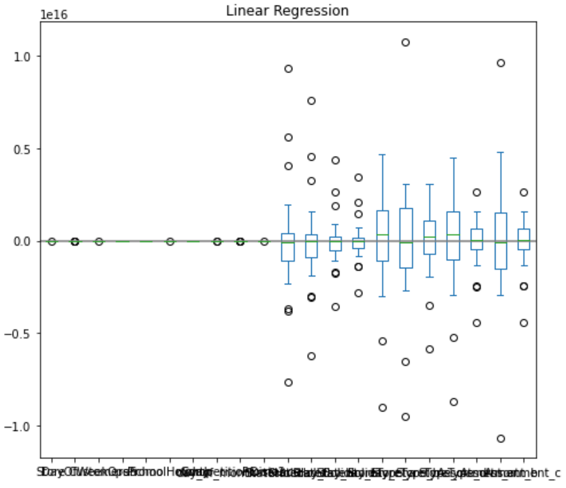
6. A summary of any modeling that was undertaken and any significant conclusions that were obtained from this analysis. Where possible do not treat models as black boxes, but endeavour to explain succinctly how the model works. (1-2 pages)

We first attempted a multivariate linear regression model to predict sales from the data. This works by calculating how linearly related the different variables are with sales and assessing by how much one increases with the other. This initial model returned an accuracy of 0.884 which is quite high however we will endeavour to see how it can be improved upon.



Before going any further with our predictive models, we created a correlation table to see how the various features within the dataset relate to one another. If our dataset contains features that correlate too closely with each other in can mean that when models are created, a change in one variable causes significant effects to another resulting in models with overly high variances. It can be seen from this graph of our initial multivariate linear regression analysis how high the variance is with our initial data.

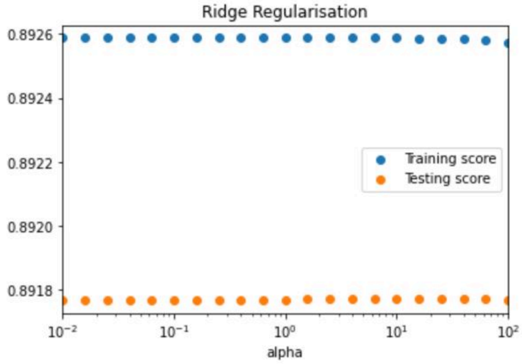




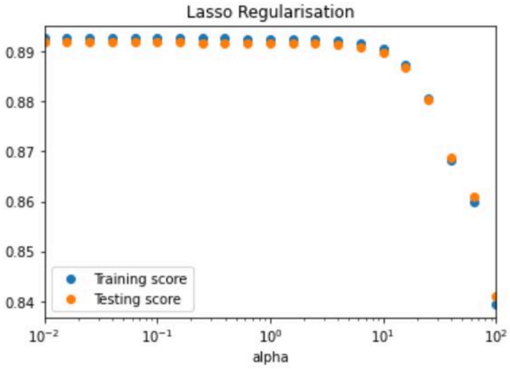
Correlation Table Variances

To help resolve these issues and ideally get more accurate predictions within our models regularisation can be employed. By regularising the data, we attempt to reduce error within our model and remove any overfitting, which is when the model or too specifically trained on just our 'train' data and can't perform well on other data, or underfitting, when the model is trained specifically enough.

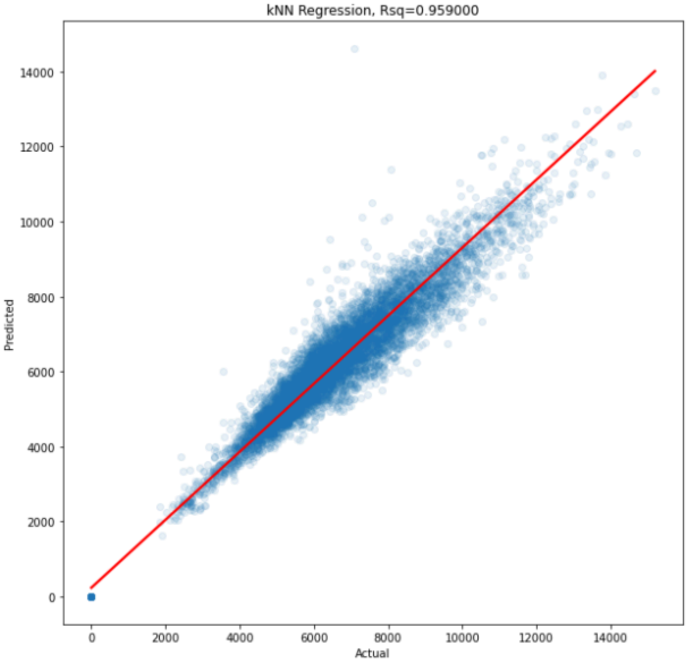
The two types of regularisation that were used to explore our data and improve the models created were Lasso and Ridge Regularisation. Lasso regularisation works by adding a penalty term for the coefficients and attempting to shrink this term down to zero. This works well when the dataset has many features with high correlations, but it also causes certain features to be lost from the data. Comparatively, ridge regularisation creates a penalty term for the variance of different features and then attempts to reduce this number to create more accurate models. This method reduces the model complexity and doesn't lose any features.



When utilised in our project lasso regularisation returned an optimal alpha of 0.001, a training score of 0.893 and a testing score of 0.884 for our multivariate linear regression. While ridge regularisation had an optimal alpha 39.811 it also returned a training score of 0.893 and testing score of 0.884 for a multivariate linear regression. Below can be seen the scores returned for different alpha values



kNN regression was also considered as a viable option for the prediction of sales. It was noted in the correlation table that sales only correlated linearly with 'customers', 'open' and 'promo'. Because of this, it was predicted that machine learning algorithms that perform well under these circumstances



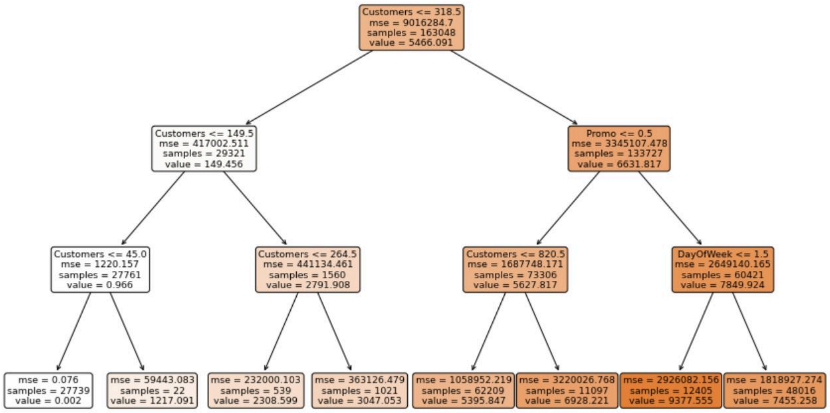
would be most useful. As kNN makes predictions based on local distance to nearby data points,linear relationships are not required. While kNN is extremely sensitive to outliers, this was not perceived as an issue as we are only working with the middle 20% of data. This ensured all extremes were removed.

When using kNN regression, different amounts of neighbours were tested, ranging from 3 to 15. After much experimentation, it was found that 5 neighbours produced the most favourable test scores and r2 scores, those being 0.95 and 0.947 respectively.

Comparatively, kNN performed very well compared to the linear algorithms. This is likely due to its ability to predict non-linear relationships with a greater accuracy

The next predictive model we created was a Decision Tree Regressor. This model works by beginning with a root node at the top of the tree. It then splits the data into two categories based on how they satisfy a certain requirement for a variable, for example X > 12. To calculate what variable it will base these decisions on it uses something called a 'Gini score' which calculates a certain score for the inequality among the values, or how well split they are, and then chooses that with the lowest score. Each node branching off from the root node then follows the same procedure until the desired depth is reached.

Decision Tree



For our initial decision tree regression, a depth of 3 was employed which returned this graph in which you can see the different branching of the nodes.

This initial model gave an accuracy of 0.861 which is quite a good performance for an initial model. We then used sklearn's GridSearchCV function which finds the best parameters for the optimal accuracy of the model. Now using a depth of 18 we were able to obtain a training score of 0.986 and a testing score of 0.950.

Following on, we then explored the creation of a Random Forest Regressor model to predict sales. A random forest regression model from the sklearn library utilises an ensemble of different decision trees and calculates the average of the predictions to find what would be the best estimate for certain values. Using 'n_estimators' = 10, which is essentially the amount of decision trees used, we were able to obtain an accuracy of 0.963 which is the best score yet ranking it as our most accurate model for predicting sales.

Additionally using this random forest regression, we were also able to obtain the feature importances showing that most important overall features are customers, day of the week and school holidays.

7. Your conclusions in relation to the original problem. (0.5-1 page)

Given the time frame the task of predicting sales six weeks in advance was achieved to a high degree with the use of random forests machine learning algorithms achieving an r squared score of 0.973. The only issues faced with this method in real life applications is customers was primarily used as a predictor which wouldn’t be available information when predicting six weeks in the future, therefore further analysis should be conducted to be able to predict customers as well. Expanding further into this project Rossmann could be encouraged to keep data on sales of specific inventory items in order to make predictions on how many of individual items or categories will sell and thus inventory can be strategically ordered to still profit and have little waste. Being able to predict the amount of customers that visit the store each day would also assist with staffing in order to schedule the optimal amount of staff to assist with customers while also reducing staffing costs.