ADS1002 Assignment 2:

ADS1002: Group Project

Introduction:

Dirk Rossmann, commonly referred to as Rossmann, is one of the largest drugstore chains in Europe. Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. The dataset contains information about store ID, day of the week, date, sales, customers, open, promo, school holiday and state holiday.

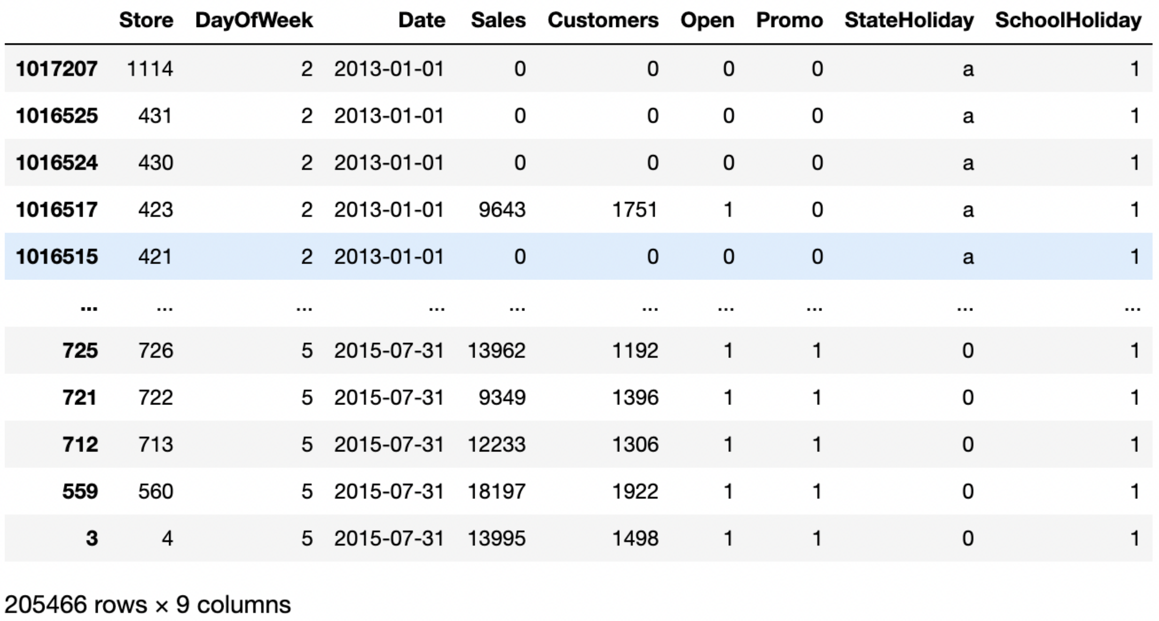
Our task:

As a group, we needed to create suitable models based on our data given to predict sales up to four weeks ahead. Our approach to this problem was as follows…

- Ryan Ngo and Yijie: cleaned and wrangled the data so that we could do useful analysis with it.

- Ben Gale: Did some summary exploratory analysis of the data so it could help us with determining the relationship between the feature variables and sales.

- Callum Battilana and Yijie: Experimented with different models with knowledge from the exploratory data analysis (prediction bounds, linear regression, Random Forests, K-Nearest Neighbours). Analysed each model to determine the best one to use.



	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday
1017207	1114	2	2013-01-01	0	0	0	0	a	1
1016525	431	2	2013-01-01	0	0	0	0	a	1
1016524	430	2	2013-01-01	0	0	0	0	a	1
1016517	423	2	2013-01-01	9643	1751	1	0	a	1
1016515	421	2	2013-01-01	0	0	0	0	a	1
...	...	...	...	...	...	...	...	...	...
725	726	5	2015-07-31	13962	1192	1	1	0	1
721	722	5	2015-07-31	9349	1396	1	1	0	1
712	713	5	2015-07-31	12233	1306	1	1	0	1
559	560	5	2015-07-31	18197	1922	1	1	0	1
3	4	5	2015-07-31	13995	1498	1	1	0	1

205466 rows × 9 columns

Data Cleaning:

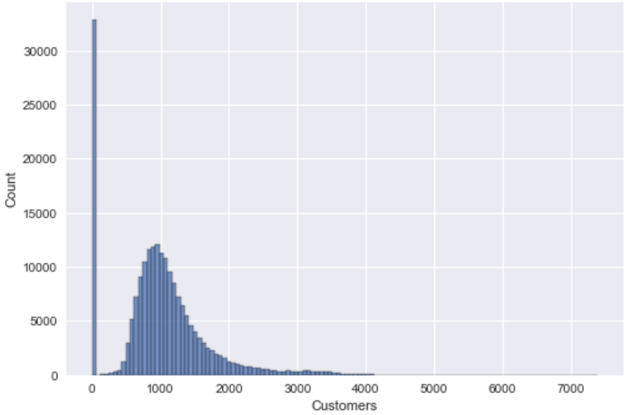
Before we begin the analysis, we first clean and filtrate the dataset. As a sample, we selected the top 20% of stores in terms of average sales for analysis. In the complete dataset, we used the average sales of each store as a criterion to screen 223 stores with the highest sales volume (top 20% stores of sales) out of a total of 1115 stores. In the data cleaning process, considering the holiday and promotion data recorded in the data corresponding to the closing day, we did not delete the data of the store closing day in the data group, and unified calculation.

After that, we converted the parameter of the 'Date' column in the dataset into a pandas datetime object to facilitate our subsequent date-related screening and analysis of the dataset.

In the final phase of data cleaning, we checked specific data to ensure smooth subsequent analysis, including whether all data was complete and whether values in the specific qualitative data were in integers.

Investigating the number of stores that are open

When we look at the frequency for the number of customers in each store, we see this…



There is an anomaly when the number of customers is 0. This is simply because about 15% of our data from our stores show they were closed at the time, and thus have no customers.

We are not going to remove the open data completely, considering the large percentage of closed data in our dataset. It is still useful to find the count of stores with certain characteristics. However, when we investigate averages of the data, we must remove the closed stores from our dataset to not distort our analyses.

We observe that there is string "0" and integer 0 in the state holiday column.

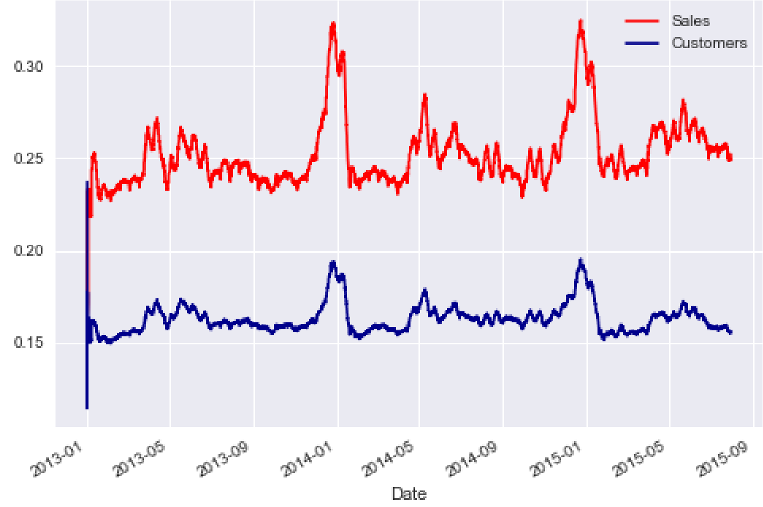
We saw that the integer zero appeared from January to March and from July to September 2014. It did not have any real reason to appear for several months just during 2014. The standard deviation of the integer zero data for all the other variables was like the string zero data, suggesting that those days did not also have missing values for any of the other variables, as there are no obvious outliers. Considering that fact, we decided to include the integer zero data with the string zero data instead of just dropping it.

Exploratory Analysis:

Our exploratory analysis was centred around which of our feature variables in our dataset we thought would have a clear influence on sales in Rossmann's stores. We justified these conclusions by real-life considerations.

Do Customers affect sales?

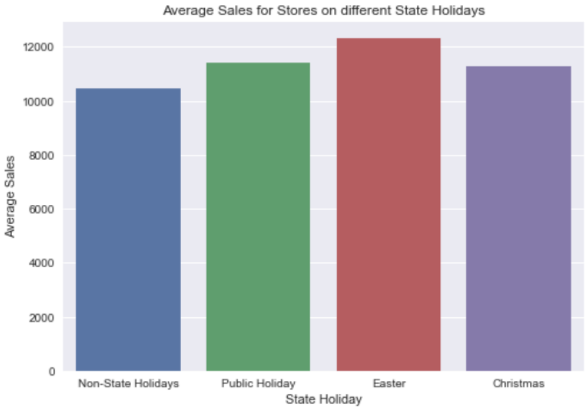
To investigate why customers are so important to affecting sales, we shall normalise the data, and compare a time series of rolling averages of sales against customers (this is so we can compare them on the same scale).

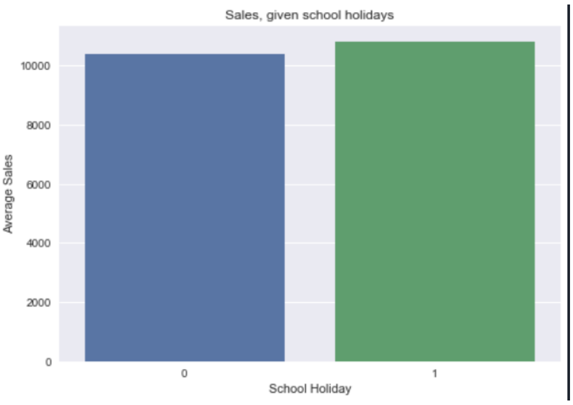


We see that these two graphs have a strong relationship between each other because the graphs have similar behaviour. The two graphs have the same shape, and they also trend the same way. When the number of customers increases per store per day, then the sales also increase. The same thing happens when there is a decrease in customers. Obviously, to make sales, one needs customers to purchase goods. So, if the number of customers changes, then sales should change by a similar amount, thus illustrating that customers do affect sales.

Impact of holidays on sales.

The impact given from holidays on sales can be shown by comparing regular days and holidays, with respect given to both state holidays and school holidays. The idea behind investigating this relationship is that holidays can potentially bring more customers into the store, as people won’t be at work, and children won’t be at school. So, for stores which are open during the school holiday period and state holidays, we assumed that they would perform better on average compared to regular days. Here are a couple of graphs to show this.

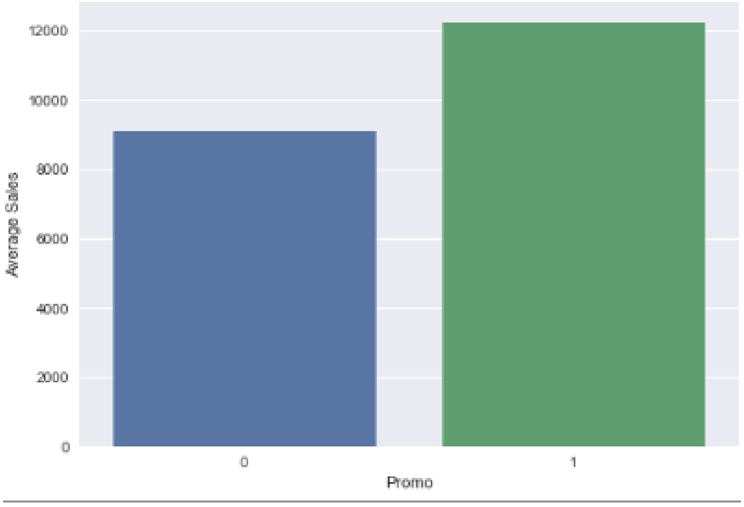




In both graphs, the blue column represents regular days. Its shown here that the regular days compared to state holidays and school holidays yield less sales per day, with state holidays providing many more sales, which shows that state holidays do have an effect on sales, however there is so much more data of regular days than state holidays, with a difference of 172,175 to 440. Although there is a massive discrepancy in data between the two categories, the data is still viable as the lack of data in state holidays doesn’t demonstrate bias in the data. Thus, state holidays have an impact on sales. However, with school holidays, we can see that there is only a minute difference in sales; a difference of only 3%. Because there isn't a large difference, we cannot conclusively say that school holidays provide a proficient enough effect of sales.

Do Promos affect sales?

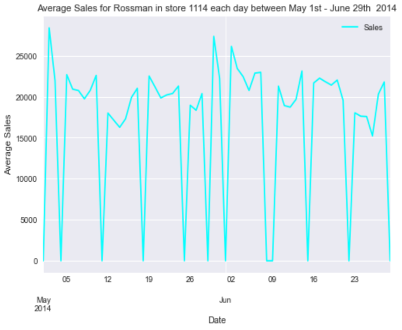
Promotions and discounts are usually used by organisations to get more people to purchase goods from a store. So, we thought that it would be a good idea to investigate whether promotions can affect sales. We graphed average sales per day for promo days and non-promo days and compared the two. The results show that the stores with promos on yielded more sales than without the promo.

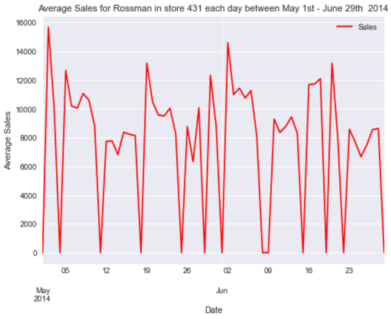


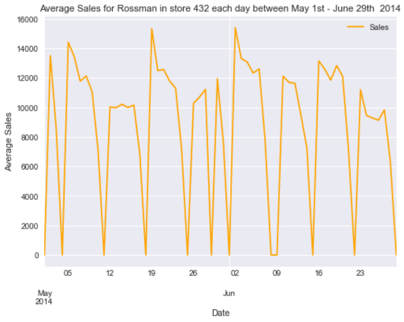
(Green is stores with promos, blue non promo)
This graph shows that promotions bring in a lot more sales per day per store. The promotions that Rossmann stores use obviously work very well.

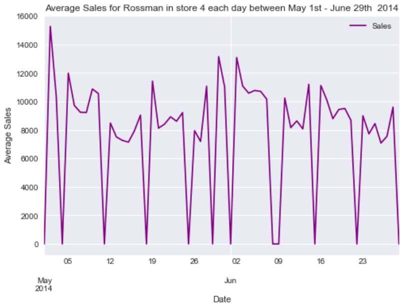
Do different days of the week bring in different amounts of sales?

Here is a time series graph of 4 different stores with respect to sales. We can see that the data has a similar trend; almost cyclic. We can see a trend where the data spikes up, decreases slightly for a few days, and then decreases massively, before increasing up the very next day

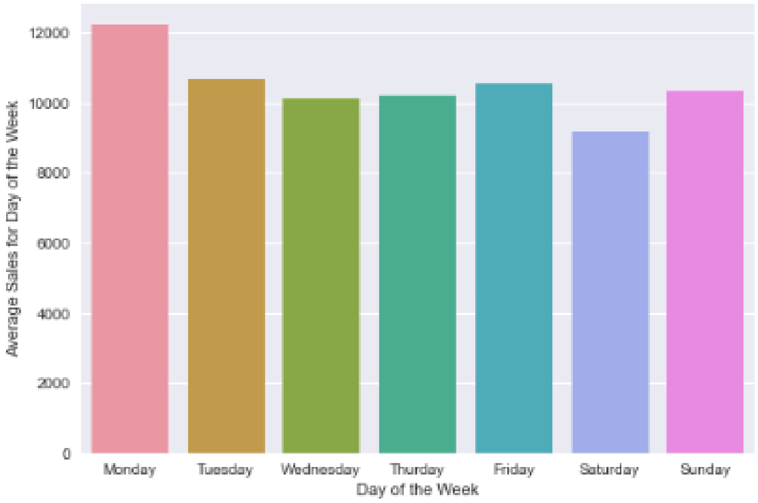








This trend is like the average weekly trend from Monday to Sunday.



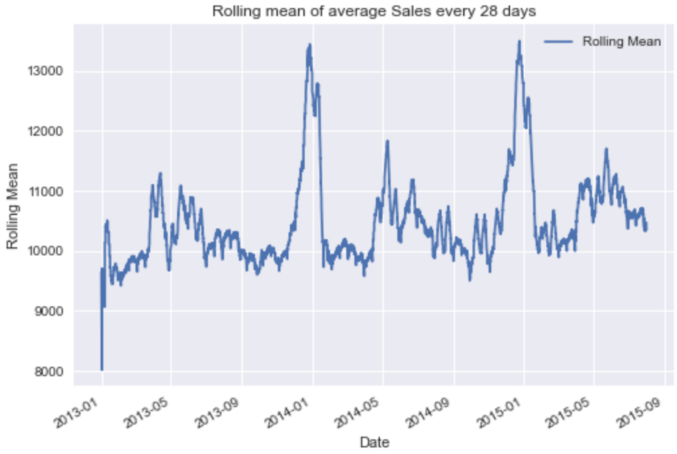
This graph shows the average sales per day of the week for stores which are open. It should be noted that stores don't regularly open on Sunday, as is common in Europe, which is why the Sunday data is quite high, as it only accounts for stores which have opened. The day of the week has a small correlation to number of sales, as Monday has a lot more average sales than other days, but all other days are similar in sales.

Forecasting Sales:

After analysing our data, we now need to create models to forecast Sales up to four weeks in advance. We will start with relatively simple models, then explore more complicated machine learning models. We will use these models to determine in different ways how the different variables affect Sales.

Model 1: Prediction Interval for Sales

In this model, we are simply finding the range of possible values that our sales could be between based on our prior data on sales.



Assuming that the standard deviation of sales over time is negligible, so that if our values are greater than this range it will not be significant. **Our simple model states that a "reasonable" estimate for average sales (if you know no prior knowledge) should be between $8020 to $13494. This is just between our maximum and minimum values for our rolling four-week daily average for sales**

Our standard deviation for the 4-week average across all stores is small (703 euros). Rossman is a multi-million-dollar company, in which we will not see such a small uncertainty per store affect its total annual revenue per store of $3.54 million (statista.com) in 2021. As such, our prior assumptions for this model are not unfounded. Furthermore, it seems like most of our average sales are clustered around a smaller range (9500-11000 euros). This means our estimate is wider than most of the sales data, which are mostly safely far from the extremes of our prediction interval.

This model could be used as a general guide to analyse how Rossman is guaranteed to earn profit. For example, if Rossman knows their average daily costs, then they can calculate how much they need to cut costs so that they are guaranteed to break even no matter what (the minimum of average daily revenue is more than their average costs). We are only forecasting four weeks ahead. We can assume that other variable costs are fixed (costs of wages, production costs), and Rossman can therefore have relatively certain predictions of how costs will change when they lower costs. This is known as *ceteris paribus.*

Model 2: Predictions based on prior knowledge

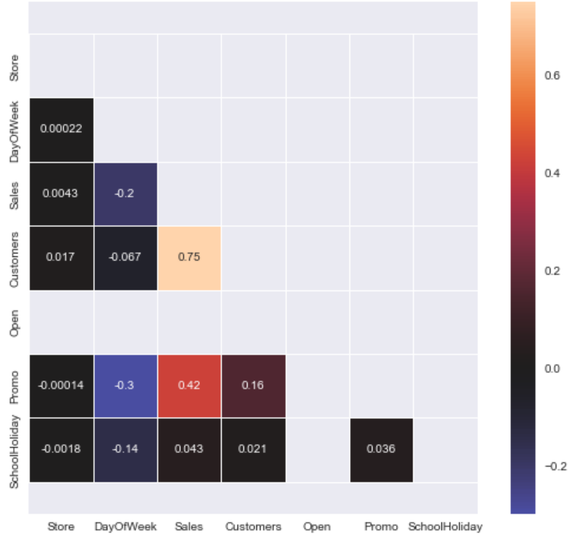
There are some glaring problems with this simple model, notably, we must assume *ceteris paribus* for Rossman's sales from 2015 to today and in the future. This is absurd, considering the many factors that could have changed Rossman's sales (expanding business, decreasing consumer spending from an inflationary period in the economy today). We believe however that our data from 2015 still can yield accurate results for today. Some of our feature variables should not have significantly changed their relationship towards Rossman's sales. For example, people are still spending a lot during Christmas time now as they were in 2013-15. We can also see that we can predict some of these features into the future (we will have four Mondays in four weeks).

We will build three machine learning models (Linear Regression, Random Forests, and K-Nearest Neighbours), which will try to predict as accurately as possible how and to what extent sales are affected by these other useful variables in our data.

Linear Regression Analysis:

Firstly, we needed to check whether linear regression was a good model for our data.

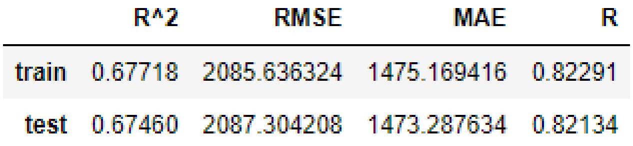
To do this, we calculated the linear correlation coefficient of each variable. We of course did not include the "open" variable, since it was obvious it would distort the model. We know if a shop is closed it does not earn any sales. Thus, we will get a too high accuracy from just predicting open is 0 to sales is 0.



We see that the linear correlation between Sales and Customers is 0.75, a **moderate positive** linear correlation. This is good enough to suggest that we can try to model a linear relationship between Customers and Sales.

Unfortunately, the correlation values for our other variables does not mean very much since they are not continuous. We therefore have to put all those variables together with customers and build a multivariate linear regression model with the target variable being Sales. We created dummy variables on the discrete data (Day of the Week, State Holiday).

Here's the error measurements for our multivariate model (our correlation coefficient is R). We made sure to normalise all the feature variables before finding these error measurements.

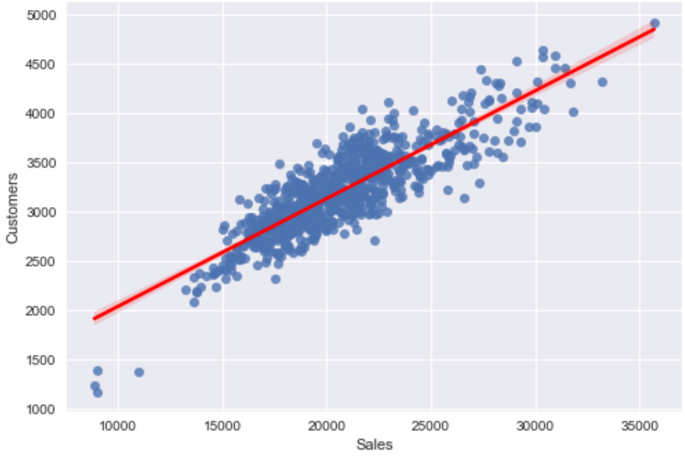


	R^2	RMSE	MAE	R
train	0.67718	2085.636324	1475.169416	0.82291
test	0.67460	2087.304208	1473.287634	0.82134

We discovered that the correlation coefficient R is 0.82 for this multivariate model test score. That is higher than the correlation coefficient for just comparing customers. Therefore, it shows that adding our other variables improves our linear regression model even more. Our Root Mean Square Error (RMSE) and our Mean Absolute Errors were 2091 and 1477 respectively. This shows that our sales were off by 2091 euros on average. Again, considering the overall revenue of Rossmann, this is a relatively small error.

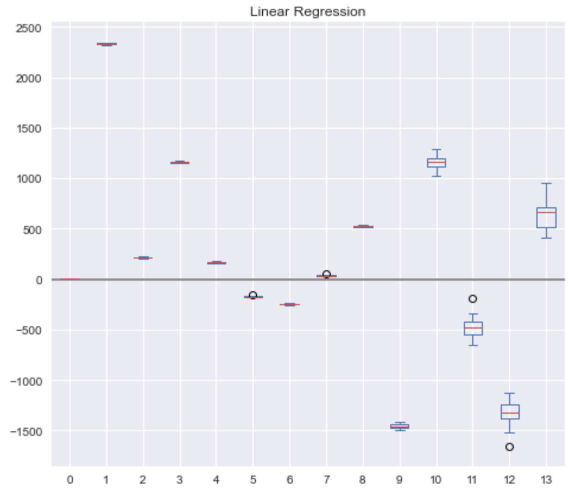
Analysing this chart, we can also recognize that this result is not (overfitted). In other words, our error measurements are about the same for when our model is tested on unknown data (test) as the data has been already trained on (train).

Let's visualise how accurately linear regression predicts a certain store (1114's) sales over customers.



In this case, our sales are scattered relatively close to the line. This shows our model should be accurate for determining how an individual store can improve Sales, based on its local changes in customers.

We will finally look at the variability of the coefficients of our linear regression model using Cross-Validation.



We can assume that the variability is negligible (small range in boxplots), meaning this model is reliable. It is also because the variables that have any visible variance have negligible importance to sales (as we will discover using Random Forest Regression).

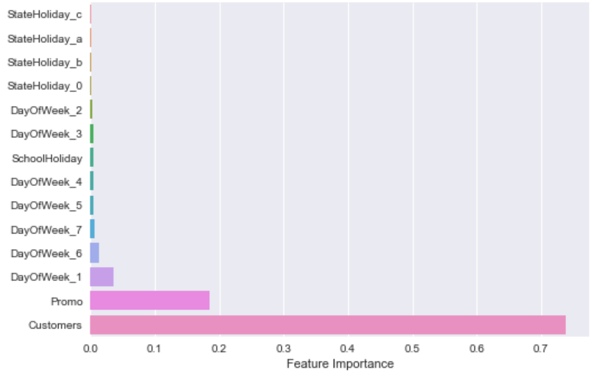
Our linear regression model is…
Sales = *4.41***(***number of customers)+2336(whether a store has promo)+211(school holiday) + 1158(Monday)+165(Tuesday)-173(Wednesday)-248(Thursday)+35(Friday)+523(Saturday)-1459(Sunday) +1160(Non-Public Holiday) - 473(PublicHoliday)-1332(Easter)+645(Christmas)*

Random Forests:

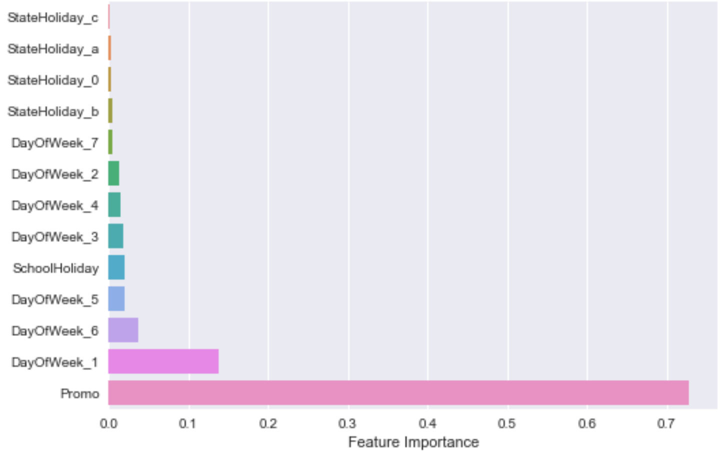
We need to compare linear regression to nonlinear models such as Random Forest Regression
Random Forests is an aggravation of random decision trees. These decision trees give weight to which variables will most affect Sales. It will then sort through the data to put it in the "boxes" with other data which is "voted" to be most likely close to that value.

Using 300 total random decision trees, we found that a max depth (number of categories filtered for data) of 10 provided the best accuracy (r squared) for our test score, as well as causing negligible overfitting to the training set (negligible difference between training and testing score). We found that the accuracy of our test score was 0.70, which in fact is slightly higher than our linear regression model r-squared score of 0.68.

Since Random Forests is the best so far, we will now use it to rank the "importance" of which feature variables (customers, day of the week ect.) impact sales the most.



We will test what happens when we remove customers for our analysis. Customers is a variable which is difficult to predict up to four weeks ahead like sales. We will compare the variables that we can predict (such a promoting items and the day of the week) with more certainty.



We now see more how promotion has a far significant impact than the day of the week. We still see that the other days of the week besides Monday do not have a real impact on sales.

Looking at this diagram, we see that Customers, promo, and whether the day is a Monday are the most important indicator of Sales according to this Random Forest algorithm. This ranking of importance is supported with most of our conclusions from our exploratory analysis. For example, we see that Monday (DayOfWeek_1) significantly changes sales than the other days of the week. Also, School Holidays do not significantly affect Sales.

We can see that if we remove the other feature variables besides customers, promo, and Monday from the model, the accuracy is still about 0.70. This shows that the other variables besides customers, promo, and Monday **do not affect sales significantly** as the accuracy is about the same as with those variables**.**

K Nearest Neighbours Regression

K Nearest Neighbours Regression predicts what the relationship is between Sales and the feature variables by calculating the average distances between k number of features which are the closest to predicting a certain target variable.

When we compute the accuracy of the model with 10 neighbours, we see that the test score is 0.66**.** Our test score accuracy for K- Nearest Neighbours is lower than all the other models.

Which model/s should be used?

Despite Random Forests giving a slightly higher accuracy than Multivariate Linear Regression, **we believe that both should be used**. A major disadvantage with Random Forests is the amount of computing power required, due to the large number of decision trees for data with large samples (Donges, 2022). Multivariate Linear Regression is a lot simpler and faster.

Multivariate Linear Regression can be used to constantly update forecast predictions based on real time data (for example, if there is a recession and customers are buying less, multivariate linear regression will update consumer demand). It is accurate enough that a general extrapolation of the sales data for up to four weeks can be feasible.

Random Forest should be used to improve the accuracy of the model itself. We could have found the best parameter for certain if we used more computational time with GridSearchcv. Random Forest can also be used to determine the importance of new feature variables that we might add to improve our model. This is useful for Rossman to determine what their business needs to change to create a substantial impact on their sales.

We should disregard K-Nearest Neighbours. It is the least accurate model, however there are other reasons too.

- **Does not work with high dimensionality:** sales are affected by many other factors besides what our data shows. The distance from Rossman's stores to its competitors, the location of where the stores are. To include all these many factors in our data, we will have to increase the dimensionality of the dataset. Increasing the dimensionality increases the number of computing steps required to calculate the minimum distance in each dimension (Soni, 2020).
- **Does not work well with large datasets:** Our dataset contains over 178,000 data points. This means that it took us a long time to calculate the accuracy of our KNN algorithm. It needs to find the minimum distance between all those 178,000 data points; therefore, it is very "costly" (Soni, 2020) in terms of time taken to compute.

Conclusion:

We have found ways to help to predict with relative accuracy how sales can change up to four weeks ahead of time for the company of Rossman. Our exploratory analysis includes that Rossman's promotional methods work to improve sales. Every Monday there was a definite cyclical increase in sales, but for the rest of the week sales are relatively the same. Meanwhile, Rossman cannot receive more sales during the school holidays. We need more data to conclude for certainty if state holidays affect sales. Customers (obviously) are strongly correlated with sales.

Our simple estimate for Rossman's average daily sales over four weeks should be between $8,000 to $13,500 based on the data we were given. We then sought to increase this accuracy by correlating important factors to sales in our that we may be able to predict will stay the same when predicting for weeks from now. We concluded that both linear and Random Forest Regression can predict accurately and efficiently **together** to correlate our variables to sales. Random Forests also allowed for us to realize that **customers, promo, and Monday** are the only three variables in our data that you need to consider for a substantial change in sales to be noticed when those variables change.

Evaluation:

Our machine learning models are still limited in predicting sales for four weeks ahead:

We realize that by far the most important variable for determining all these models is customers (due to its overwhelming importance to the other variables on the Random Forests algorithm). Customers is of course difficult to be predicted very accurately four weeks ahead, while the other data is controllable (the day of week we are certain to predict). If we tried our machine learning models to predict quantitively sales with more certain data, our models would be useless due to only having discrete data. We can only tell the importance of those other variables to affect sales with Random Forests. Knowing the importance of the non-customer variables does not allow us to predict sales **quantitatively** four weeks ahead, however. This is what is useful for a business to know most of all. We therefore decided to keep sales, since we wanted to try and quantify the prediction of sales over four weeks.

References

Donges, N. (2022). *Random Forest Algorithms: A Complete Guide | Built In*. Builtin.com. https://builtin.com/data-science/random-forest-algorithm#procon

Soni, A. (2020, July 3). *Advantages And Disadvantages of KNN*. Medium. https://medium.com/@anuuz.soni/advantages-and-disadvantages-of-knn-ee06599b9336

Rossman. (2022). *ROSSMANN | Unsere Marken*. ROSSMANN - Corporate Website. https://unternehmen.rossmann.de/ueber-uns/unsere-marken.html

Rawat, K. (2020, October 31). *Rossmann Store Sales Prediction*. Analytics Vidhya. https://medium.com/analytics-vidhya/rossmann-store-sales-prediction-998161027abf#c67