UDATE REPORT: DATA PRE-PROCESSING

Introduction

As of 2021, Rossmann continues to dominate the drugstore chain industry in Germany (Statisa, 2024), undoubtedly implying that in a rapidly dynamic world of business, mitigating risks and capitalising opportunities are important factors in a successful business (Trkman, 2010). However, sales fluctuate based on many key factors like seasonality and locality – evident in the finding that retail revenues during holiday seasons show a 16.6% growth rate (Statista, 2019b), and consumer behaviours also differ during promotion seasons like Black Friday (Trkman, 2010). Whilst this variation could be a challenge in the market, a huge retail company like Rossman can opportunistically consider these factors by accurately forecasting sales to circumvent overproduction (Islam & Amin, 2020), maximise profits by predicting costs (Huang et al., 2019; Isleam et al., 2021), and in turn, improve customer service. Sales forecasting can be done by observing the overall or particular products of the company's sales patterns. Nevertheless, predictions if done inaccurately may carry significant consequences to businesses. For instance, Apple in 2013 dropped in stock values due to inaccurate forecasting of the iPhone 5C model which led to overstocking and profit loss (Guardian, 2013). Hence, it is imperative that sales prediction methods are strategically improved for better accuracy and efficiency - this can be done by handling historical datasets appropriately. This report therefore aimed to use past data from Rossmann to predict 6 weeks of daily sales as accurately as possible through emphasising on data pre-processing of the dataset and exploring major data preparation tasks.

Methodology

Exploratory Data Analysis

Three datasets were investigated for sales forecasting which were 'stores.csv' - consisting information on 1,115 stores, as well as 'train.csv' and 'test.csv' showing historical sales data in two different timepoints of 1/1/2013-31/7/2015 and 1/8/2015-17/9/2015 respectively. Understanding the general insights of the dataset allows for determining the approaches to data pre-processing - and as the aim was to carry out sales predictions with accuracy, we justified the use of our selections in preparing data and forecasting to satisfy this. We used the *pandas* module in Python to carry out data exploration. Figures 1-3 summarises statistics of the datasets.

	Store	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear
count	1115.00000	1112.000000	761.000000	761.000000	1115.000000	571.000000	571.000000
mean	558.00000	5404.901079	7.224704	2008.668857	0.512108	23.595447	2011.763573
std	322.01708	7663.174720	3.212348	6.195983	0.500078	14.141984	1.674935
min	1.00000	20.000000	1.000000	1900.000000	0.000000	1.000000	2009.000000
25%	279.50000	717.500000	4.000000	2006.000000	0.000000	13.000000	2011.000000
50%	558.00000	2325.000000	8.000000	2010.000000	1.000000	22.000000	2012.000000
75%	836.50000	6882.500000	10.000000	2013.000000	1.000000	37.000000	2013.000000
max	1115.00000	75860.000000	12.000000	2015.000000	1.000000	50.000000	2015.000000

Figure 1: Summary statistics for 'stores.csv'

	Store	DayOfWeek	Sales	Customers	Open	Promo	SchoolHoliday
count	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06
mean	5.584297e+02	3.998341e+00	5.773819e+03	6.331459e+02	8.301067e-01	3.815145e-01	1.786467e-01
std	3.219087e+02	1.997391e+00	3.849926e+03	4.644117e+02	3.755392e-01	4.857586e-01	3.830564e-01
min	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	2.800000e+02	2.000000e+00	3.727000e+03	4.050000e+02	1.000000e+00	0.000000e+00	0.000000e+00
50%	5.580000e+02	4.000000e+00	5.744000e+03	6.090000e+02	1.000000e+00	0.000000e+00	0.000000e+00
75%	8.380000e+02	6.000000e+00	7.856000e+03	8.370000e+02	1.000000e+00	1.000000e+00	0.000000e+00
max	1.115000e+03	7.000000e+00	4.155100e+04	7.388000e+03	1.000000e+00	1.000000e+00	1.000000e+00

Figure 2: Summary statistics for ‘train.csv’

	Store	DayOfWeek	Open	Promo	SchoolHoliday
count	41088.000000	41088.000000	41077.000000	41088.000000	41088.000000
mean	555.899533	3.979167	0.854322	0.395833	0.443487
std	320.274496	2.015481	0.352787	0.489035	0.496802
min	1.000000	1.000000	0.000000	0.000000	0.000000
25%	279.750000	2.000000	1.000000	0.000000	0.000000
50%	553.500000	4.000000	1.000000	0.000000	0.000000
75%	832.250000	6.000000	1.000000	1.000000	1.000000
max	1115.000000	7.000000	1.000000	1.000000	1.000000

Figure 3: Summary statistics for ‘test.csv’

- **Quality**
  - ‘train.csv’ was robust as there were no missing values, giving more confidence about the analytical integrity. However, ‘stores.csv’ and ‘test.csv’ had unknown values in certain attributes which were handled and explained in the next sections.
  - Data quality is further established by the absence of duplicates in all three datasets. Outliers present in Sales and Customers particularly may suggest factors like special promotions which should be taken into account.
- **Relevant Variables**
  - In terms of variables, the relevant attributes to observing sales were mainly Sales and Customers against factors like promotion, holidays, and day. Observing the interactions of these variables are helpful in understanding the relevance and influence to sales predictions.
  - Figure 4 shows the pairplot of variables and deduces that;
    - Scatterplots generally show that the more the customers, the higher the sales in a store which is sensible.
    - The negative correlation between the days over the week and Sales/Customers indicate that sales and customers tend to slightly reduce by the end of the week.
    - Promotions when observed with Sales demonstrate that when there is one, sales and customers tend to increase.

- School holidays did not show a correlation as strong as promotional season but maintained a moderate increase in sales/customers.

- Figure 5 shows a sample of a temporal analysis from the ‘train.csv’ dataset and generally shows a spike in sales and customers at the end of the year potentially due to holiday season and school holidays.

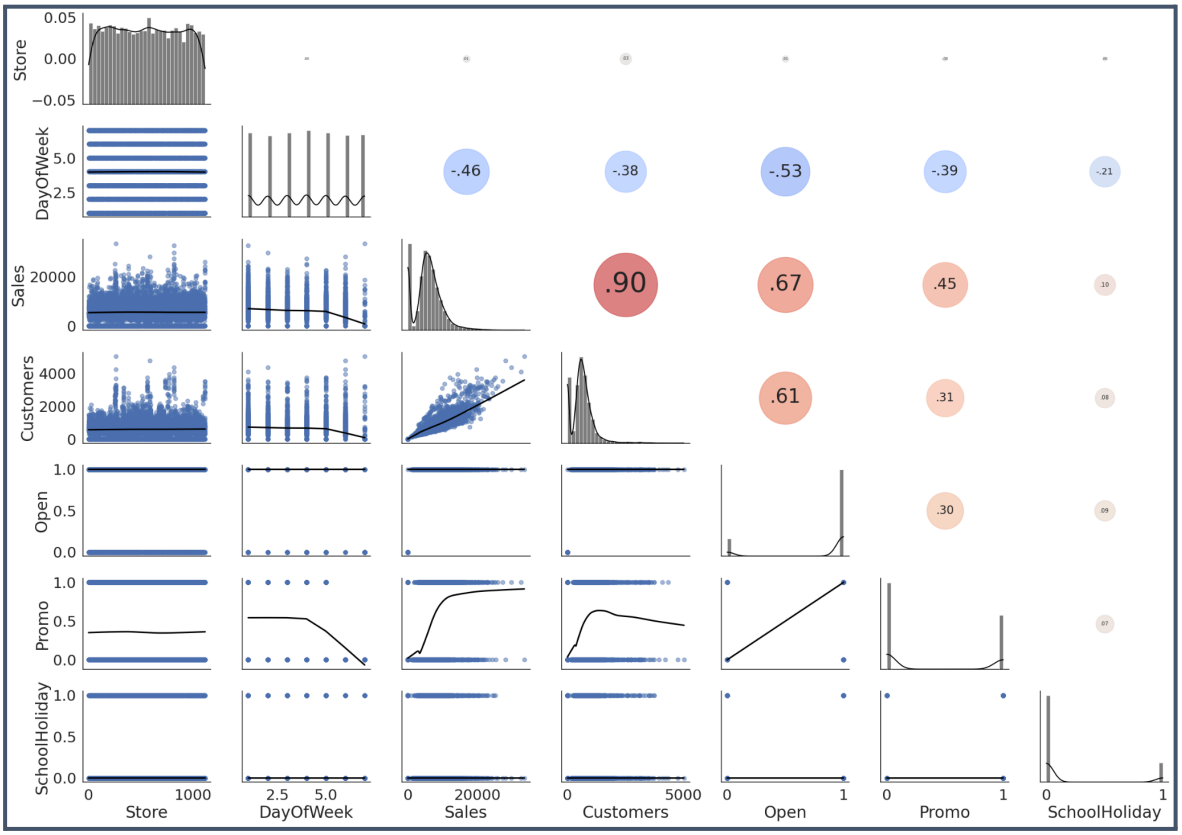

Figure 4: Pairplots of Variables

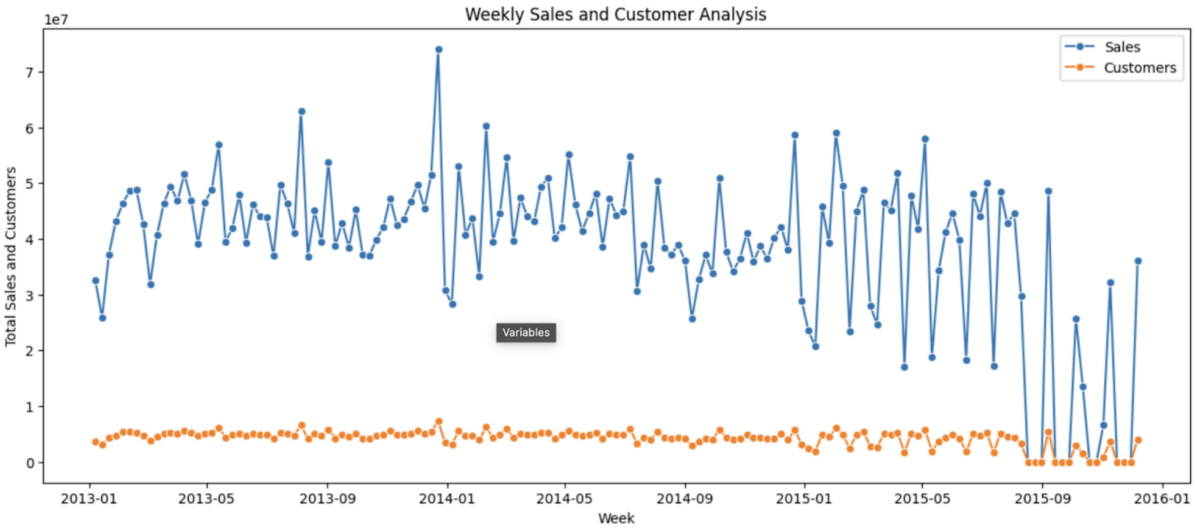

Figure 5: Weekly Sales/Customer Analysis

- ***Data Preparation***

Data Cleaning and Transformation

Using the LabelEncoder from the *sklearn* library, data transformation was done by encoding categorical variables into numerical variables. This step was crucial to prepare data for improved further use in machine learning modelling (see Figure 5).

```
StoreType                 91319 non-null  int64
Assortment                91319 non-null  int64
PromoInterval             91319 non-null  int64
Season                    91319 non-null  int64
```

Figure 5: Labels encoded into numerical

In terms of completeness, the missing values present in datasets implied that the incompleteness had to be handled accordingly to reduce noise. In terms of accuracy, visualisations showed that there were outliers present in data. Regardless, we decided not to remove these values due to the nature of the data being historical and therefore needs to be preserved where possible. Lastly, the consistency of the dataset was satisfactory as it conforms to the expected values.

Data Integration

To further maintain the consistency throughout training and testing the model, we linked datasets 'train.csv' with 'stores.csv' and 'test.csv' with 'stores.csv' through the unique attribute of 'Store' using an inner join as shown in Figure 6.

```
final_train = pd.merge(train, stores, on='Store', how='inner')
final_train
```

```
final_test = pd.merge(test, stores, on='Store', how='inner')
final_test
```

Figure 6: Dataset linkage

Handling Missing Data

There were no missing values for 'train.csv'. However, in the 'test.csv' dataset, *Sales* and *Customers* information were missing which is expected as these were supposed to be forecasted using historical data from 'train.csv'. Considering the nature of missing data, the 11 values unknown in *Open* were missing at random and handled by imputing with mode as this gives an idea of whether other stores usually open on the date or not.

In the 'stores.csv' dataset, the missing data in *CompetitionDistance* (3 values), *CompetitionOpenSinceMonth* and *CompetitionOpenSinceYear* (354 values each) were also missing completely at random - these were information regarding the nearest competitor store. To handle this, we imputed these numerical values with median instead because measures like mean would have been more sensitive to outliers which can influence skewness. This was considered the most appropriate way to handle it assuming that other pharmaceutical stores would have similar openings.

The 544 missing data on *Promo2SinceWeek*, *Promo2SinceYear*, and *PromoInterval* were not missing at random. Referring to Promo2, stores with these missing values simply did not participate in promotion (=0). Therefore, these values were dropped.

```
Store                         0
StoreType                     0
Assortment                    0
CompetitionDistance           3
CompetitionOpenSinceMonth   354
CompetitionOpenSinceYear    354
Promo2                        0
Promo2SinceWeek             544
Promo2SinceYear             544
PromoInterval               544
dtype: int64
```

Figure 7: Missing value counts

- ***Data reduction***

To gain a more comprehensible insight into the datasets and forecast sales more accurately, some feature engineering was considered by creating new indicators based on temporal features and customer-related features (see Figure 8).

- Temporal features that were utilised were dividing data into dates (month, year, weekends), seasons (winter, spring, summer, fall), and lag features (SalesLag1-7 and CustomersLag1-7).
- Customer-related features used were Sales per Customer and Average Sales per Month to give a clearer insight into customer satisfaction and profitability.

is_weekend	Season	SalesLag1	SalesLag2	CustomersLag1	CustomersLag2	SalesLag7	CustomersLag7	SalesPerCustomer
0	2	0.0	0.0	0.0	0.0	0.0	0.0	9.482883
0	2	5263.0	0.0	555.0	0.0	0.0	0.0	9.194139
0	2	5020.0	0.0	546.0	0.0	0.0	0.0	9.143403

Figure 8: Temporal and Customer-related Features

In terms of feature selection, variables that were most highly correlated were selected for model training as they were most relevant to sales prediction.

	SalesLag1	SalesLag7	CustomersLag1	CustomersLag7	SalesPerCustomer
Sales	0.294356	0.655547	0.321644	0.665564	0.202968
Customers	0.273915	0.659143	0.486636	0.844641	-0.341676

Results

Time-series forecasting using Seasonal Autoregressive Integrated Moving Average (SARIMA) was trained on the cleaned dataset without the unknown 6 weeks from test.csv. Similar to ARIMA, SARIMA was chosen due to the consideration of its seasonal factor relevant to our dataset. An advantage of using this over other models is because it has been successfully utilised in retail forecasting (Goh & Law, 2002) similar to Rossman's data. However, due to its linear approach, SARIMA may not be able to grasp nonlinear trends

which is quite important in sales (Ensafi et al., 2022). The correlation matrix done on the train dataset showed that ‘Sales’ and ‘Customers’ had the highest correlation with sales/customer lag. Therefore, these were tested with SARIMA. The accuracy of the model performance was evaluated using metrics like Mean Square Error (MSE), and coefficient of determination (R^2). Commonly utilised in model assessments, these metrics give an understanding of evaluation given that MSE is to assess accuracy, MAE for robustness and R^2 to understand the amount of variances on data. Although typically utilised, it is important to take note of other non-negligible influences like outlier presence and the ML model’s implicit assumptions which can impact model efficiency. In theory, a smaller difference between the MSEs of the train and test sets, the higher the model efficiency. Higher values of R^2 also indicate a better model fit.

Conclusion, implications, recommendation

In summary, the three datasets on historical sales and stores by Rossmann were investigated to predict 6-weeks of sales through an exploratory data analysis. To ensure higher accuracy, we initially assessed the data quality and variables relevant to sales forecasting. An emphasis was placed on the data preparation tasks where data were cleaned, transformed, linked and visualised appropriately. We also considered feature reductions to produce more coherent insights by adding temporal and customer-related features. Using pre-processed features that were most highly correlated, we then modelled the data using time series modelling (SARIMA) and trained it. The performance was evaluated using appropriate metrics.

Our approach on understanding the key factors which influence Rossmann’s sales is in line with current literature in the retail industry that generally agrees that this is comprehended by analysing sales trends, with seasonality and promotions playing significant impacts in the patterns (Ensafi et al., 2022). With feature engineering, the new indicators we created based on season and time provided usefulness in the model training as this approach was also demonstrated in other sales forecasting reports (see Huang et al., 2022).

However, many limitations were encountered in this forecast as there is generally no best approach to comprehend issues in time-series sales predictions (Zhang & Kline, 2007). Firstly, various other factors may impact sales but the dataset was limited by only the ones provided like promotion and competitors. This could be addressed by conducting an exploration of other factors like macroeconomic information. Moreover, missing data at random posed a significant limitation which could be mitigated by extending historical data to provide a more accurate assumption and reliability on imputation. Also, due to the SARIMA method’s inability to grasp fluctuating trends, Ensafi et al. (2022) in their comparative study on modelling in sales forecasting recommended considering a Seasonal Support Vector Regression approach to mitigate the seasonality issue.

1643 words

REFERENCES

Dominic R. (2013). Apple profits and shares fall amid fears over iPhone 5c sales. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2013/oct/28/apple-shares-fall-results-iphone-5c.

Ensafi, Y., Amin, S. H., Zhang, G., & Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning–A comparative analysis. *International Journal of Information Management Data Insights*, 2(1), 100058.

Goh, C., & Law, R. (2002). Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention. *Tourism management,* 23(5), 499-510.

Huang, T., Fildes, R., & Soopramanien, D. (2019). Forecasting retailer product sales in the presence of structural change. *European Journal of Operational Research*, 279(2), 459-470.

Islam, S., & Amin, S. H. (2020). Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques. *Journal of Big Data*, 7, 1-22.

Trkman, P. (2010). The critical success factors of business process management. *International Journal of Information Management,* 30(2), 125-134.

Islam, S., Amin, S. H., & Wardley, L. J. (2021). Machine learning and optimization models for supplier selection and order allocation planning. *International Journal of Production Economics,* 242, 108315.

Koptyug, E. (2024) Drugstore chains: Stores in Germany 2021, Statista. (Accessed: 24 February 2024).

Zhang, G. P., & Kline, D. M. (2007). Quarterly time-series forecasting with neural networks. *IEEE transactions on neural networks,* 18(6), 1800-1814.