Report

Content

Introduction

In the context of the retail industry, the accurate prediction of sales is of paramount importance to retailers[1]. However, this is a task that is fraught with difficulty. This project takes a major German pharmacy chain as the research object, with the aim of predicting the daily sales of 1,115 pharmacies for the next six weeks, by exploratory data analysis[2], Feature engineering[3] includes processes such as data pre-processing[4], handling of missing data, and feature transformation, etc., after which model training is performed and the performance.

This report contains and will explain in detail the exploratory data analysis, data engineering – data preprocessing, handling missing data and data integration, and the use of machine learning models for prediction[5].

Method

In Feature Engineering section, missing values are handled by introducing new variables[6], and technical coding is used for categorical variables using label encoding[7] to make the data easier to understand for the model. Variables are extracted from date to capture the seasonal trend[8] of the data.

In the model section, a random regression model that can handle high-dimensional data and outliers is selected based on the EDA results, and a time series model is tried to capture the time dependence in the data. In addition, 5-fold cross-validation, R^2 and RMSPE are used to evaluate the performance of the model.

[1] Robert Fildes, Shaohui Ma, Stephan Kolassa (Retail forecasting: Research and practice, 5/12/2019, https://doi.org/10.1016/j.ijforecast.2019.06.004)
[2] IBM (https://www.ibm.com/think/topics/exploratory-data-analysis)
[3] Harshil Patel (29/4/2024, Feature Engineering Explained | Built In)
[4] Idan Novogroder (30/4/2024, https://lakefs.io/blog/data-preprocessing-in-machine-learning/)
[5] towardsanalytic (January 16, 2023, 9 Top Machine Learning Algorithms for Predictive Modeling)
[6] Statisticseasily (What is: New Variable - Understanding Its Importance)
[7] Ajitesh Kumar (8/2/2024, Sklearn LabelEncoder Example - Single & Multiple Columns)
[8] Jason Brownlee (15/8/2020, How to Identify and Remove Seasonality from Time Series Data with Python - MachineLearningMastery.com)

Feature Engineering

0	Store	1115	non-null	int64
1	StoreType	1115	non-null	object
2	Assortment	1115	non-null	object
3	CompetitionDistance	1112	non-null	float64
4	CompetitionOpenSinceMonth	761	non-null	float64
5	CompetitionOpenSinceYear	761	non-null	float64
6	Promo2	1115	non-null	int64
7	Promo2SinceWeek	571	non-null	float64
8	Promo2SinceYear	571	non-null	float64
9	PromoInterval	571	non-null	object

Table 1 Information of store data

Table1 presents a comprehensive overview of the data, including the quantity of non-empty data, the data type, the necessity of reviewing existing data prior to performing data preprocessing, and the significance of data analysis.

A thorough analysis of the project's textual description reveals that the store type symbolizes distinct store models, the category denotes the classification level of the store, the data encompasses records related to competition, information about discounts is recorded in the promo records, and records concerning holidays are documented in the holiday records. This initialization establishes the foundation for subsequent processing of missing values.

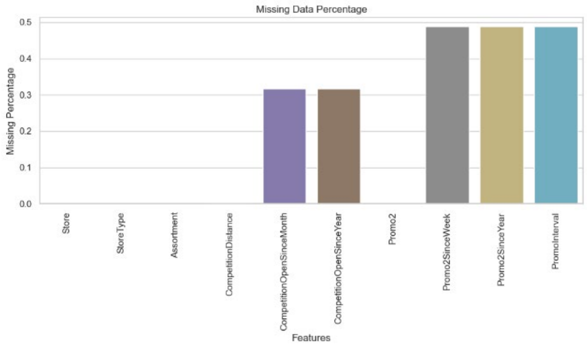


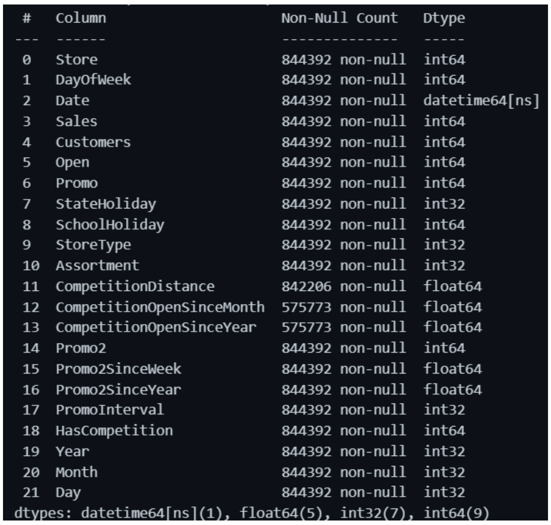
Figure 1 Distribution of Missing Value

The utilization of visual analytics for the identification of missing data within the store environment are observed in competing stores and promotional information. However, based on the previous analysis, it can be found that the Promo-related information is empty because the store is not in a promotional state, which can be inferred from Promo2. The competition store-related information and the competition store distance are NULL in same time, which is thought to be because there is no competition store. It is imperative to note that the arbitrary allocation of values may introduce noise and resulting in biased outcomes.

Store	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	HasCompetition
291	d	a	nan	nan	nan	0
622	a	c	nan	nan	nan	0
879	d	a	nan	nan	nan	0

Table 2 new variable - HasCompetition

Consequently, a novel variable, designated as "HasCompetition," has been incorporated to denote the presence of competing stores within the commercial environment. Furthermore, the presence of a missing value in the train data, accompanied by an open of 0, signifies that the store was not opened. This missing value is deemed essential for non-business operations and assign a value of 0 to these missing values[9].



```
 #   Column                     Non-Null Count   Dtype
---  ------                     --------------   -----
 0   Store                      844392 non-null  int64
 1   DayOfWeek                  844392 non-null  int64
 2   Date                       844392 non-null  datetime64[ns]
 3   Sales                      844392 non-null  int64
 4   Customers                  844392 non-null  int64
 5   Open                       844392 non-null  int64
 6   Promo                      844392 non-null  int64
 7   StateHoliday               844392 non-null  int32
 8   SchoolHoliday              844392 non-null  int64
 9   StoreType                  844392 non-null  int32
 10  Assortment                 844392 non-null  int32
 11  CompetitionDistance        842206 non-null  float64
 12  CompetitionOpenSinceMonth  575773 non-null  float64
 13  CompetitionOpenSinceYear   575773 non-null  float64
 14  Promo2                     844392 non-null  int64
 15  Promo2SinceWeek            844392 non-null  float64
 16  Promo2SinceYear            844392 non-null  float64
 17  PromoInterval              844392 non-null  int32
 18  HasCompetition             844392 non-null  int64
 19  Year                       844392 non-null  int32
 20  Month                      844392 non-null  int32
 21  Day                        844392 non-null  int32
dtypes: datetime64[ns](1), float64(5), int32(7), int64(9)
```

[9] Learn statistics easily (What is: Zero-Fill - LEARN STATISTICS EASILY)

Figure 2: Data information summary

In addition, to facilitate subsequent model training, the datetime format is converted for date conversion, and Label Encoding is performed on the type variables to convert this data into a type that is understandable by machine learning algorithms[10]. And combine the store information with the sales information in the data table by stores'ID. In addition to this, considering that there may be seasonal or other cyclical patterns[11] of change, create Features: year, month, day.

EDA (Exploratory data analysis)

Single Variable

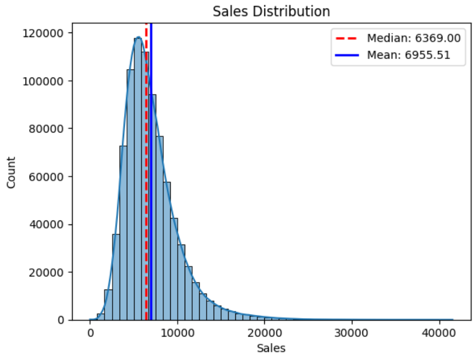


Figure 3 distribution of Sales in training set

Since our focus is on sales, we focus on sales-related information.

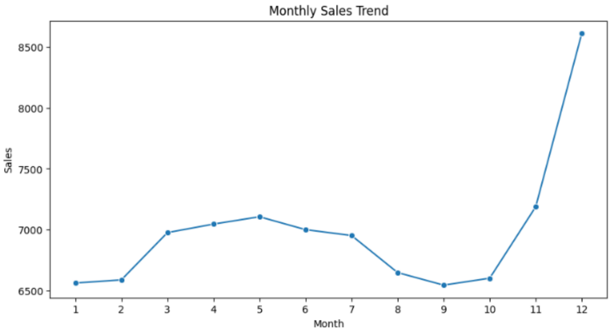


Figure 4: Monthly Sales Trend

Considering the seasonal cyclic trends, a rough visualisation of Month shows that sales from January to August are relatively stable with small fluctuations, and sales increase rapidly at the end of the year in November and December. This may relate to

[10] HogoNext (31/10/2024, How to Data Conversion for Machine Learning - Preparing Your Data for AI Algorithms - HogoNext)

[11] Rob J Hyndman (14/12/2011, Cyclic and seasonal time series – Rob J Hyndman)

seasonal factors, such as the Christmas holiday.

Multiple Variable

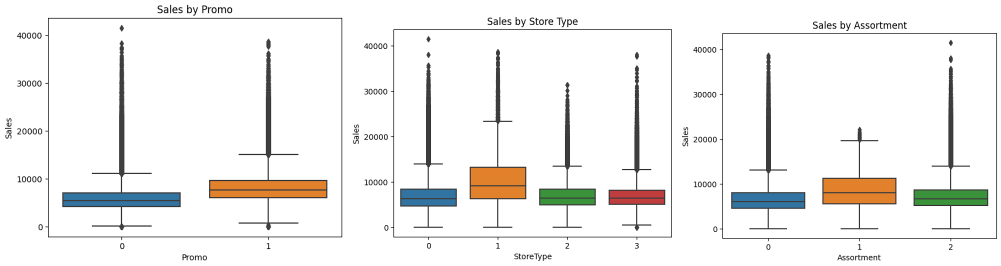


Figure 5: Sales in Promo, Store Type and Assortment

It can be seen that the participating promotion has more sales than the non-participating promotion in the concentrated data. However, the extreme values are not as large as the non-participating promotion, and the sales are more widely distributed in the non-participating promotion. It can be seen that promotions have a certain impact on sales. Sales are also better in stores with StoreType b.

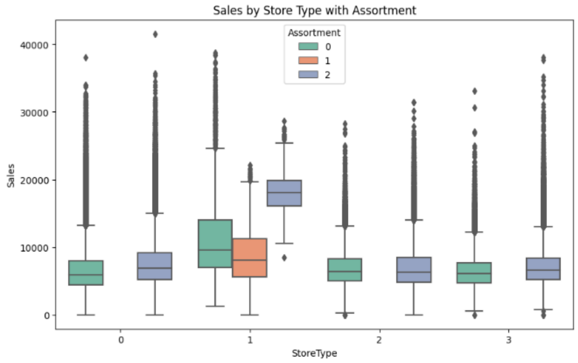


Figure 6: Sales with Stroe Type and Assortment

It can be seen that when the assortment is “extended” and the store type is b, the sales are extra high, and no matter what the assortment is, the sales of store type b are higher than those of other store types.

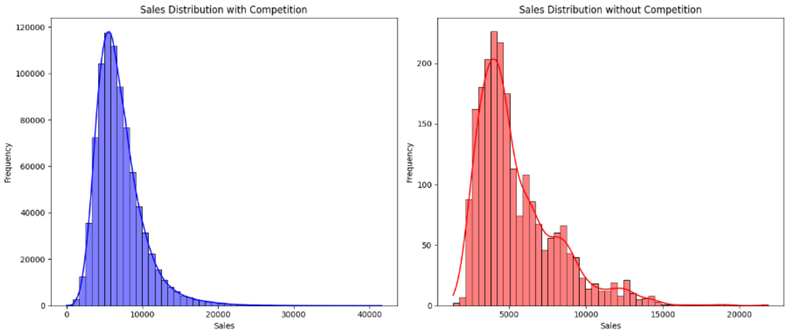


Figure 7: Sales distribution between competition and non-competition

Stores with competitors tend to have lower sales, but occasionally very high sales. Stores without competitors have bimodal sales data, which means that there may be two main customer groups or sales models. In contrast, where there is a competitor, the peak in sales is higher, but the overall distribution is also more concentrated, suggesting that competition can lead to polarisation. There are also cases of unusually high sales regardless of the presence of competing stores. This is a bit different from what I expected at first. At first, I thought that no competitor could generate higher sales, but this idea ignores the fact that a lively area is more attractive to shops, which leads to the existence of competing shops. However, sparsely populated places do not have commercial appeal and naturally do not have competing shops.

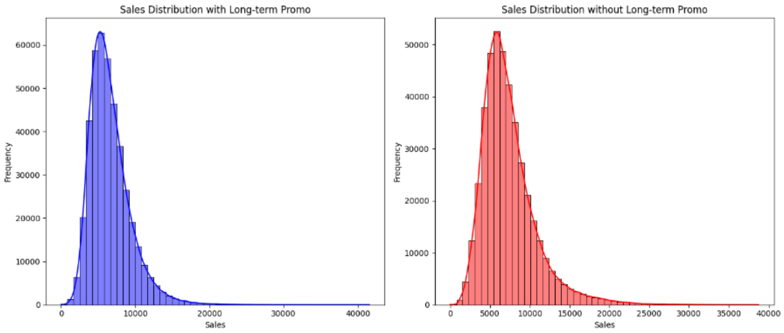


Figure 8: Sales between long-term and non-long-term

It can be seen that the sales of long-term promotions have significantly higher peaks and the overall distribution is also skewed to the right, meaning that long-term promotions can effectively increase sales.

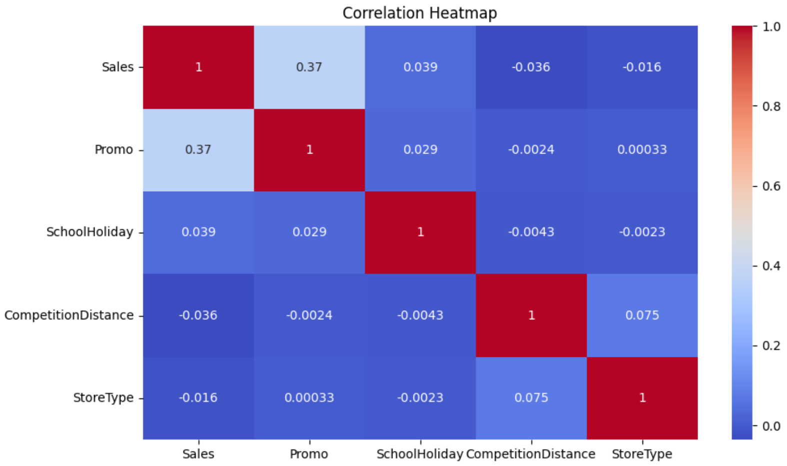


Figure 9: Correlation Heatmap

It can be seen that promo and the distance between the competitor has the greatest impact on sales.

Time Series Analysis

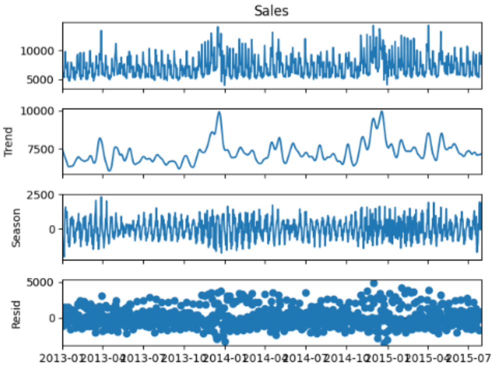


Figure 10: Sales in Trend, Seasonality and Residuals

The figure shows that sales from 2013 to mid-2015 fluctuated significantly in a cyclical pattern, with a cycle length of about one year and a general growth trend towards the end of the year. It is worth noting that after removing the trend and seasonality[12], there are still some random fluctuations in the data, indicating that there are some uncaptured factors or outliers in the data.

Model

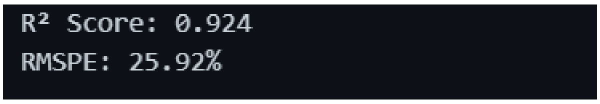
Selection

After the previous exploratory data analysis, it can be concluded that there are a large number of outliers in the data set. It is necessary to select a method that is tolerant of outliers and has good capabilities for processing high dimensional and categorical data. As the visualisation has shown that time, promotions, competitors and store type all have an impact on sales, the training data will be high-dimensional if these variables are retained. Finally, the data appear to be in a non-linear relationship.

A random forest regression[13] model was chosen because it is robust to outliers and missing values and can handle categorical variables (random forests also have classification models). Importantly, it performs well and is stable with high-dimensional data.

[12] Leon Yen (13/11/2023, What is Time Series Analysis? Definition, Types, and Examples)
[13] Nima Beheshti (2/3/2022, https://towardsdatascience.com/random-forest-regression-5f605132d19d)

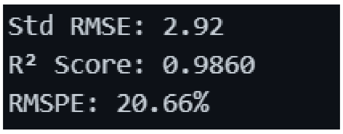
Assessment



```
R² Score: 0.924
RMSPE: 25.92%
```

Figure 11: R^2 Score and RMSPE

These two indicators show that the model is performing relatively well. The R^2 value[14] is close to 1, indicating that it can explain the variability of the data well. The RMSPE[15] is relatively low, indicating that the predicted value of the model is close to the actual value.



```
Std RMSE: 2.92
R² Score: 0.9860
RMSPE: 20.66%
```

Figure 12: 5fold-cv

Reduce stochasticity and transition fitting and improve the stability and generalisation ability of model performance estimates by using a 5-fold cross-validation[16] to partition the data multiple times to calculate the model's performance across different data partitioning.

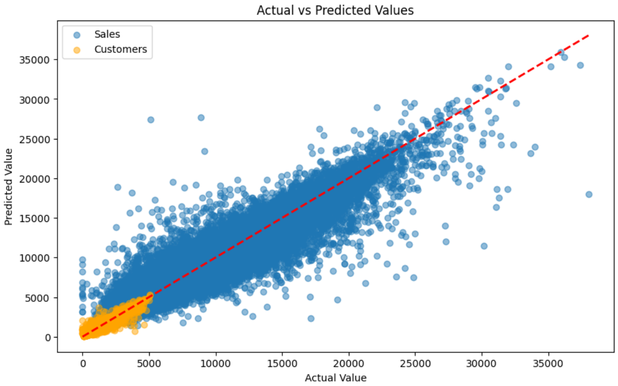


Figure 13: real value vs predicted value

The blue dots are concentrated near the red dotted line, which means that the model's prediction of sales is close to the actual value. The fact that both variables are more or less equally distributed on either side means that there may be random errors[17].

[14] Natalia Afek (Coefficient of determination R^2: what is it and how to interpret it? - Predictive Solutions)

[15] Sudeept Singh Yadav (2022, Root Mean Square Error of Prediction - an overview | ScienceDirect Topics)

[16] Geeksforgeeks (30/12/2024, Cross Validation in Machine Learning - GeeksforGeeks)

[17] Anne Helmenstine (26/7/2021, Systematic vs Random Error - Differences and Examples)

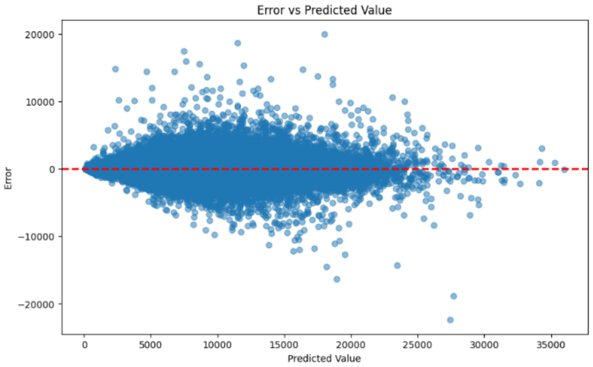


Figure 14: Residual Plot

Most of the points are concentrated near the red line and are more evenly distributed, indicating that the model fits relatively well. The distribution of the errors is relatively random, again proving the previous argument that the errors are random and points very far from the line may be outliers.[18]

In addition, since the previous analysis revealed the existence of periodic fluctuations related to time, a time series model was trained.

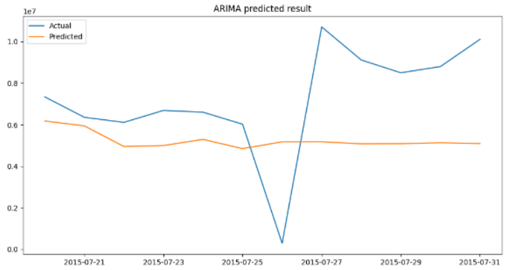


Figure 15: ARIMA predicted result

The model has an R² Score of 0.986 and an RMSPE of 20.66%. Overall, the model can track the trend of actual values well most of the time, but there are significant errors at certain points. Judging from the picture, it is not performing well when dealing with sudden increases and decreases. This shows that the model performs well in capturing the long-term trend of data, but it is not very accurate in the short term when there are drastic fluctuations.

Prediction

Use the model to predict the test data set

[18] Zach Bobbitt (17/2/2023, What is Considered a Good vs. Bad Residual Plot?)

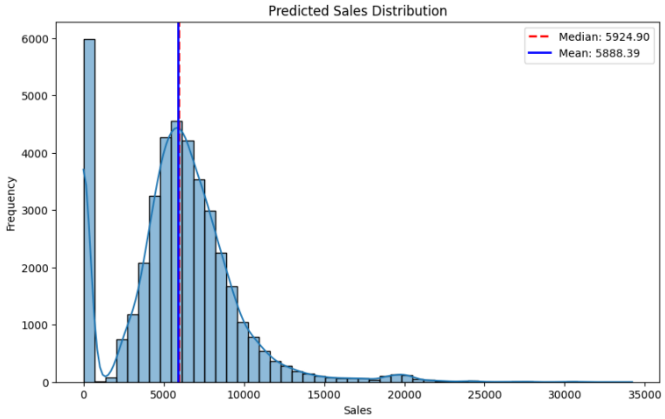


Figure 16: Predicted Sales distribution

Compared with the sales distribution of the training set, it can be seen that the two distributions tend to be right-skewed, with similar trends, and the median and mean are also located in similar positions. This indicates that the predicted results are roughly similar to the original data trend. However, the simulation of some extreme positions is not very ideal.

Conclusion

This study aims to predict daily sales of 1,115 German pharmacies using exploratory research, feature engineering and machine learning. Key findings include that competitor proximity, promotions and StoreType (especially b) significantly affect sales. In high-traffic areas, stores with an extended assortment show higher sales. In addition, there is a clear seasonal time trend, with the peak during the holiday season (November-December).

The random forest model trained on this prediction achieved good results (**R^2 = 0.98**, **RMSPE = 20.66%, Std RMSE = 2.92**), although extreme sales remain challenging.

Based on the recommendations from the previous analysis, the retailer should prioritize location analysis (e.g. to avoid clusters of competitors) and promotional strategies to maximize sales. Long-term promotions show a positive effect on variance, suggesting that phased activity can balance customer engagement and margins. In addition, consider increasing the number of stores in the extended category to meet customer demand for this type of store. Plan inventory and promotions in advance in November and December (the holiday season) to maximize sales.

Finally, it is worth considering the limitations of the project: there is a risk of bias in the treatment of missing values and there is no proper treatment of extreme values. Although a model that is not sensitive to outliers was chosen, the cause of the extreme values was not investigated in depth, which may lead to the loss of important information. In addition, the dataset lacks statistics on customer information, which limits the insights into purchasing behaviour, and does not include external factors such as weather and unexpected events.

Appendix