PROJECT SUMMARY

📖 Problem Statement

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied. Two datasets are given: one with store data and the other with historical sales data of 1115 stores from January 2013 to July 2015. The main objective is to understand existing data and after identifying the key factors that will predict future sales, a predictive model will be built for making forecasts about future sales.

📖 Approach

1. Understanding the business task.
2. Import relevant libraries and define useful functions.
3. Reading data from files given.
4. Data pre-processing, which involves inspection of both datasets and data cleaning.
5. Exploratory data analysis, to find which factors affect sales and how they affect it.
6. Feature engineering, to prepare data for modelling.
7. Modelling data and comparing the models to find out most suitable one for forecasting.
8. Conclusion and recommendations to boost sales.

📖 Exploratory Data Analysis

The following insights were gained from EDA:

- There were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week.
- Store type B though being few in number had the highest sales average. The reasons include all three kinds of assortments specially assortment level b which is only available at type b stores and being open on sundays as well.

- The outliers in the dataset showed justifiable behaviour. The outliers were either of store type b or had promotion going on which increased sales.
- Store type B was open on all seven days of the week and had more sales than any other store type and promotion had a positive effect across all store types.

📖 Modelling

	Model_Name	Train_MAE	Train_MSE	Train_RMSE	Train_R2	Train_Adj_R2	Test_MAE	Test_MSE	Test_RMSE	Test_R2	Test_Adj_R2
0	Decision Tree Regressor	0.00000	0.000000	0.000047	1.000000	1.000000	0.014203	0.000466	0.021580	0.915750	0.915700
1	Random Forest Regressor	0.00304	0.000022	0.004640	0.996143	0.996143	0.010328	0.000245	0.015653	0.955673	0.955647
0	Random Forest Tuned	0.00304	0.000021	0.004622	0.996173	0.996173	0.010342	0.000244	0.015617	0.955878	0.955852

📓 Conclusion

Businesses use sales forecasts to determine what revenue they will be generating in a particular timespan to empower themselves with powerful and strategic business plans. Important decisions such as budgets, hiring, incentives, goals, acquisitions and various other growth plans are affected by the revenue the company is going to make in the coming months and for these plans to be as effective as they are planned to be it is important for these forecasts to also be as good.

The work here forecasts the sales of the various Rossmann stores across Europe for the recent six weeks and compares the results from the models developed with the actual sales values.

Some important conclusions drawn from the analysis are as follows:

- there were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week. This validates the hypothesis about this feature.
- The positive effect of promotion on Customers and Sales is observable.
- Most stores have competition distance within the range of 0 to 10 kms and had more sales than stores far away, probably indicating competition in busy locations vs remote locations.
- Store type B though being few in number had the highest sales average. The reasons include all three kinds of assortments specially assortment level b which is only available at type b stores and being open on Sundays as well.

- The outliers in the dataset showed justifiable behaviour. The outliers were either of store type b or had promotion going on which increased sales.
- Random Forest Tuned Model gave the best results and only 0.021% improvement was seen from the basic random forest model which indicates that all the trends and patterns that could be captured by these models without overfitting were done and maximum level of performance achievable by the model was achieved.

Recommendations

- More stores should be encouraged for promotion.
- Store type B should be increased in number.
- There's a seasonality involved, hence the stores should be encouraged to promote and take advantage of the holidays.
-

References

Andrew Udell, 'Predicting E-Commerce Sales with Random Forest'. [Online].

Available: https://towardsdatascience.com/predicting-e-commerce-sales-with-a-random-forest-regression-3f3c8783e49b

Builtin.com, 'Random Forest'. [Online].

Available: https://builtin.com/data-science/random-forest-algorithm

Machine Learning Mastery, 'Random Forest for Time Series Prediction'. [Online].

Available: https://machinelearningmastery.com/random-forest-for-time-series-forecasting/