DOI: https://doi.org/10.61762/pjstvol8iss1art003

PLATFORM - A Journal of Science & Technology

Received: 7 April 2025, Revised: 23 June 2025, Accepted: 26 June 2025, Published: 30 June 2025, Publisher: UTP Press,

DATA ANALYTICS PREDICTIVE MODEL TO PROMOTE UP-SELLING AND CROSS-SELLING ACTIVITIES BASED ON CONSUMER'S BUYING PATTERN

Kian San Yap, Tieng Wei Koh

Department of Computer and Information Science, Universiti Teknologi PETRONAS, Malaysia

*E-mail: yap_18001106@utp.edu.my

ABSTRACT

In the current competitive market, companies need to utilise data-driven approaches to maximise sales and improve customer interaction. This research intends to create a predictive analytics and recommendation system to enhance up-selling and cross-selling efforts for dealers. By employing machine learning models such as linear regression, random forest, and extreme gradient boosting (XGBoost), the system predicts product demand, facilitating improved inventory control and sales enhancement. Moreover, an association rule mining (Apriori algorithm) method is utilised to uncover frequent item associations, producing customised product suggestions to increase transaction values. Research results showed that the most suitable prediction method is XGBoost for this research case, with an accuracy of $R^2 = 0.810543$. The research solution is proven to be scalable through data analytics using three different synthetic datasets from Kaggle, including Rossmann Store Sales, Bakery Transaction Data, and Corporación Favorita Store Sales. The research integrates these elements into an interactive Power BI dashboard, providing dealers instant insights into customer purchase patterns, sales outcomes, and inventory trends. This research provides a one-stop solution for dealers to observe sales trends, and with insights into recommended items, dealers can achieve up-selling and cross-selling. Key performance indicators (KPI) are shown in the Power BI dashboard as a centralised display, thus dealers can make decisions to promote products to customers with a higher success rate. In future, the research will hopefully expand data sources to industrial data, improve model performance, and integrate more advanced technology into the Dealer Management Systems (DMS). The DMS developed is scalable and can deal with expansion and upscale upgrades in future work.

Keywords: *Predictive analytics, recommendation system, linear regression, random forest, XGBoost, association rule mining, Apriori algorithm*

INTRODUCTION

With the advancement and revolutions of technology and the rapid growth of the retail and distribution size, the demand for data-driven insights to have a clear glance at the company's sales and marketing strategies is increasing. By leveraging the study on customer purchasing behaviour, sales trend, up-selling and cross-selling strategies, and sales product bundling, this research introduces a one-stop solution as a Dealer Management System (DMS). With this DMS, dealers can have a centralised analysis dashboard that consolidates data science tools, including machine learning predictive models, and data mining recommendation systems. As the dealers promote their product to the customer, the DMS operates as a shopping assistant that provides suitable recommendations that can be bundled or upsized, achieving a higher transaction value. Since most traditional retail or dealer systems are not capable of integrating the use of recommendation systems, this research aims to eliminate this weakness by creating a powerful DMS. From the DMS insights, dealers can examine customer buying patterns, forecast upcoming sales, and create personalised product suggestions.

RELATED WORK

In large enterprises with many outlets, dealers, and distributors, a centralised DMS is vital for the management team to have quick and precise insights into the current business status. For example, famous e-commerce companies such as Shopee and Lazada implemented sophisticated predictive and recommendation algorithms to achieve upselling and cross-selling. Their applications capture user preferences, stay time on certain products and product interactions to perform predictions and recommend suitable products. The precision of recommendations contributes to higher sales and customer retention [1]. The DMS solution developed in this research aims to serve as an industrial-ready solution for the sales and marketing field using data-driven approaches.

By accommodating a DMS with predictive analytics and recommendation systems, companies can organise different campaigns from time to time to motivate dealers to get more sales with various rewarding systems. Based on the insights provided by DMS, dealers can leverage the data-driven tools to get more sales by upselling and cross-selling. Distributors can also trace the sales trend based on location with higher demand and provide suitable product supplies and further increase company sales [2].

The three machine learning predictive models studied include linear regression, followed by random forest, and extreme gradient boosting (XGBoost). Linear regression is usually used to show the relationship between a dependent variable and one or more independent variables by sketching the best-fit line [3]. On the other hand, random forest combines multiple decision trees to form an ensemble learning technique to make predictions [4]. This technique is commonly used to deal with complex and non-linear relationships. Another method involved is the XGBoost; researchers utilised XGBoost to forecast, showcasing its effectiveness in dealing with massive datasets [5].

In this research, the model evaluation plan is to build three predictive models with different methods, then compare their performance side by side to choose the best model to be implemented in the DMS. The chosen predictive model will be integrated seamlessly into the dashboard together with the recommendation systems [6].

The recommendation system will be implemented using association rule mining and the Apriori algorithm. Apriori algorithms are a method used for association rule learning, which identifies patterns in products that are frequently purchased at once by recognising regular itemsets and formulating association rules [7]. For instance, the algorithm will put product B on the recommendation of Product A if Product A is usually bought together with Product B [8]. These association rules enable the system to propose suitable recommendations.

METHODS

The flowchart in Figure 1 shows the processes and research direction for data sourcing, model development, and dashboard development.

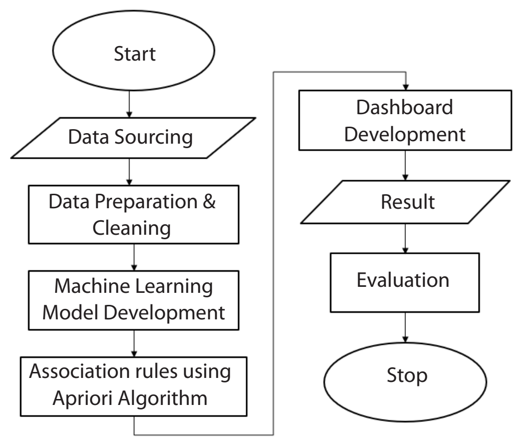

Figure 1 Research flow

Data Preparation and Data Processing

At this stage, the data source for this research is online datasets from Kaggle. Three datasets have been shortlisted, including Rossmann Store Sales, Bakery Transaction Data, and Corporación Favorita Store Sales.

The Rossmann Store Sales dataset utilises sales information from more than 3,000 European pharmacies to predict demand influenced by seasonal trends, holidays, and promotional events. Machine learning models, such as XGBoost, random forest, and linear regression, will enhance upselling and cross-selling tactics while optimising stock allocation.

Even with discrepancies such as absent transaction IDs, the Bakery Transaction Data collection monitors consumer buying behaviour. Following data cleansing, association rule mining will uncover commonly purchased product pairs. By incorporating this information into a DMS, customer satisfaction will increase, and personalised recommendations will be enhanced.

The Corporación Favorita dataset offers crucial transactional, promotional, and geographic information to enhance sales strategies and manage inventory. Predictive algorithms will evaluate seasonal changes, marketing strategies, and past sales patterns. Store clustering will assist in customising marketing strategies, and insights about holidays and events can aid in predicting demand changes for improved inventory management and focused advertising.

Data Analytics

As the baseline predictive model for this research, linear regression is used to analyse the relationship between daily sales and influence factors such as promotions and competitors' locations. The baseline model uses statistical methods such as bar charts and correlation heatmaps to visualise the relationship between a dependent variable and one or more independent variables [9]. A prototype predictive model is built using linear regression with the Rossmann Store Sales dataset. The predictive model saves a portion of the data as unseen data and tests the model's ability to make predictions with the trained model. This prediction mimics demand forecasting, where the model is implemented on industrial data to capture and analyse real-time data.

Random forest model combines multiple decision trees to form an ensemble learning model to make predictions. By adjusting necessary parameters, the model development is repeated using Random Forest to perform demand forecasts. The runtime of the computation is longer compared to linear regression since this model is more computationally intensive. The model development is repeated using XGBoost.

Finally, the model was redeveloped using XGBoost, with hyperparameter tuning applied. Due to the algorithm's complexity and intensive computations, the process required extended runtime and multiple processors to enable parallel processing. XGBoost is well-suited for handling missing values and large datasets, offering high predictive accuracy.

When dealing with a larger dataset like Corporación Favorita Store Sales, frequent itemset mining is studied, and calculations using support factors include confidence, lift, and support. Exploratory data analysis, feature engineering, scaling, log-transform, hyperparameter tuning, and fine-tuning are needed to ensure the computation is efficient and uses all processors in hardware. The models developed are scalable, so the code can be applied to different datasets and the different parameters. Then, the computation results are recorded and discussed later.

Recommendation Systems

The recommendation system is developed using association rules mining and the Apriori algorithm, where the algorithm determines the frequent itemsets and builds up corresponding rules. This system is easy to interpret since the rules are straightforward and provide actionable insights into item associations, which fit well with this research goals of recommending suitable items to achieve upselling and cross-selling.
The algorithm works well with transactional data and provides excellent transactional analysis. However, the drawbacks include that many excessively low-value rules are generated when the system loops through the dataset [10]. The computation of the Apriori algorithm is intensive on large datasets; the development execution runtime is long when the algorithm is applied to the Bakery Transaction datasets to build a prototype.
The association rules obtained are exported and cleaned according to the performance metrics, support, confidence, and lift. The rules are tested by selecting different kinds of bakery goods, and the recommendations are shown.

The developed algorithm was subsequently applied to the Corporación Favorita Store Sales dataset to perform association rules mining. Due to the dataset's significantly larger size compared to the other two, a vast number of rules were generated, resulting in a cluttered and unmanageable rule set. To address this, the rules were processed by evaluating their confidence levels, followed by a cleaning step to refine and retain only the most relevant associations.

Power BI Dashboard

The Power BI dashboard serves as the framework for the DMS. The DMS integrates both predictive models and recommendation systems output and provides actionable insights at the dashboard [11]. The predictive model will visualise the sales trend, and a specific region can be focused on with the location and outlet filter while the analysis is being done. The recommendation systems will operate dynamically such that when dealers choose one item, there will be four more recommended items. The Power BI dashboard also includes key performance indicators (KPI) such as total revenue, best-performing outlet, and best-selling items. To sum up, the Power BI dashboard consolidates all the components needed to build the DMS by providing a dynamic and interactive interface.

RESULT AND DISCUSSION

Data analytics is performed by feeding data to three different predictive models: linear regression, random forest, and XGBoost. Each model has its strengths and drawbacks. The evaluation results of each model are analysed, and the three models are compared, revealing the most suitable model for this DMS.

Machine Learning Models

Figure 2 shows the actual compared to predicted values for a linear regression model. Linear regression assumes a linear relationship between the features and the target variable, aiming to fit a straight line that minimises the overall error. This graph demonstrates a significant spread of data points around the red dashed line (representing the ideal predictions where actual values align with predicted values). The model struggles to comprehend complex relationships, particularly with higher actual sales numbers, resulting in an underprediction of outcomes. The presence of outliers and deviations from the diagonal line suggests that the linear model may not have the necessary flexibility to address all disparities in the dataset.

Figure 3 shows the efficiency of a random forest model, an ensemble method that utilises multiple decision trees to make predictions. Compared to linear regression, we notice a more dispersed and clustered pattern of predictions. The model seems to perform well with average actual sales data but struggles with outliers, leading to overfitting or skewness in certain regions. The horizontal clusters of predictions indicate that the model tends to predict a certain range of values, a common issue with decision tree models when managing continuous outputs. This results in a tiered or stepped look rather than a seamless linear development.

Figure 4 displays the predictions generated by a sophisticated gradient boosting method known as XGBoost, which constructs trees sequentially to reduce errors. XGBoost's predictions align more closely with

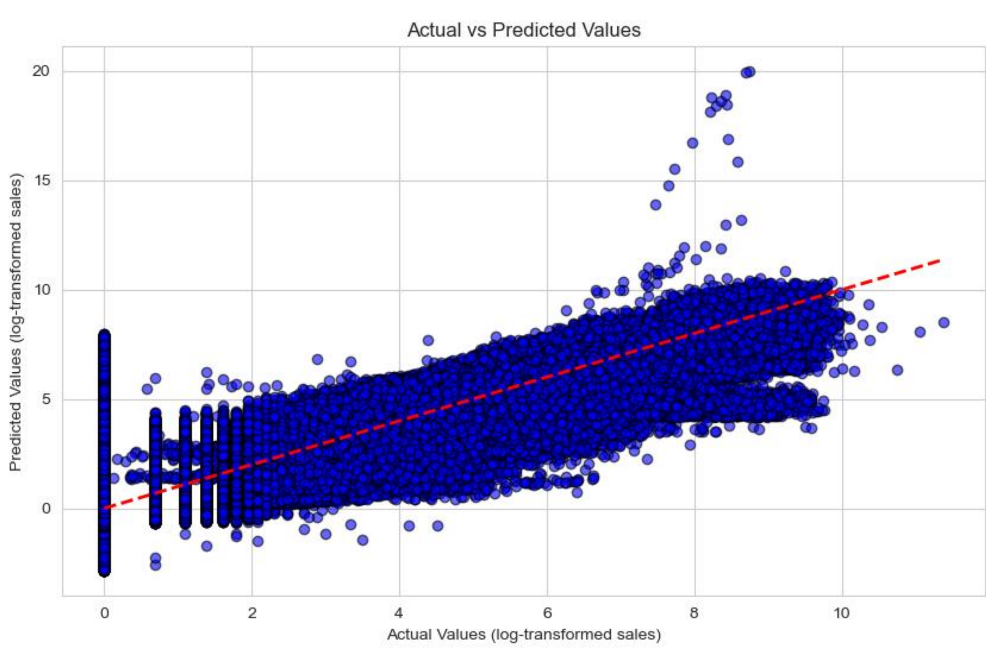

Figure 2 Model prediction using a linear regression model

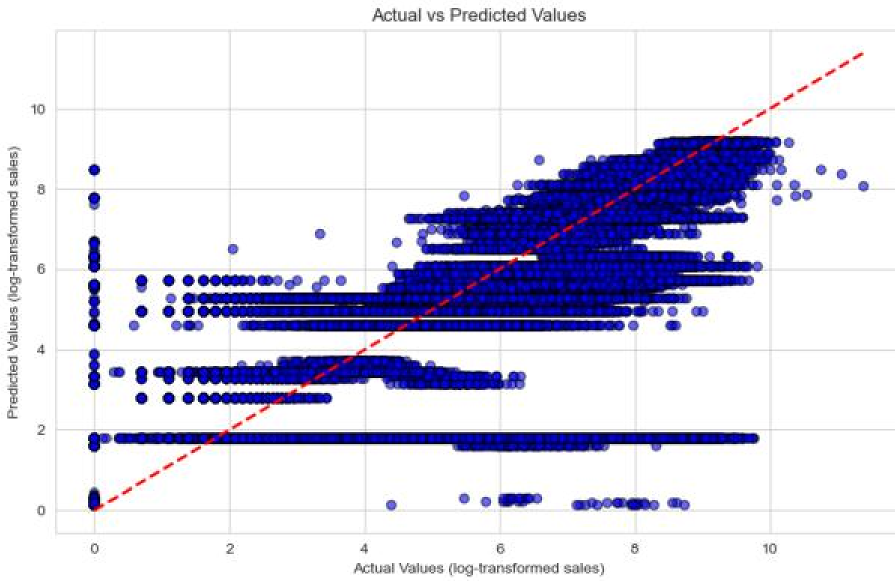

Figure 3 Model prediction using random forest model

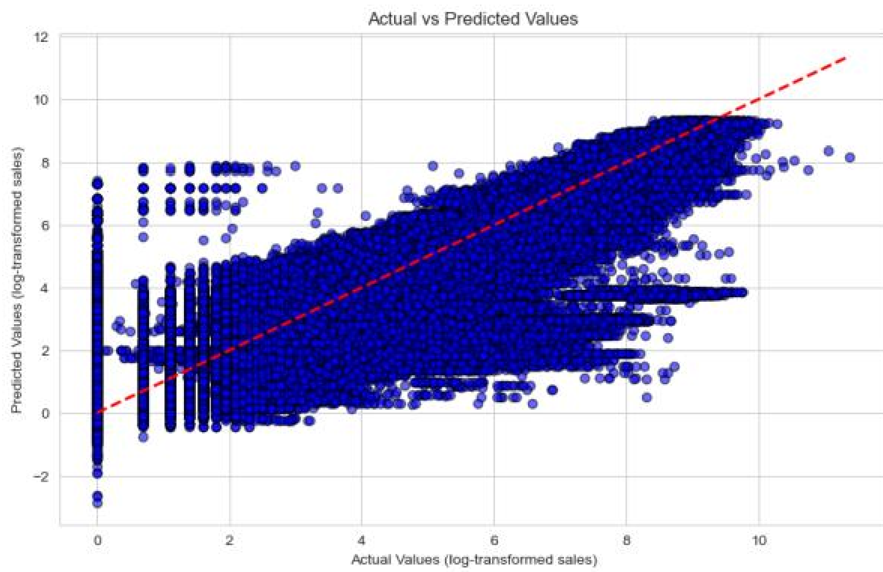

Figure 4 Model prediction using XGBoost model

the actual values compared to the Random Forest model, as indicated by a distribution nearer to the red diagonal line. This indicates that XGBoost exhibits better generalisation when identifying more intricate patterns within the data. Compared to the other models, the point distribution appears more fluid and smoother, suggesting enhanced prediction accuracy and reduced bias.

To evaluate the effectiveness of the machine learning models, four main assessment metrics were considered: Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 (coefficient of determination). The outcomes from the assessment are shown in the table above.

Table 1 Evaluation matrix summary

Model	MSE	MAE	RMSE	R^2
Linear Regression	2.145099	1.041979	1.464616	0.705005
Random Forest	3.608123	1.517053	1.899506	0.503809
XGBoost	1.377665	0.85685	1.17374	0.810543

MSE assesses the average of the squared differences between actual values and predicted values. A smaller MSE signifies greater predictive precision.
MSE is computed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (1)$$

Based on the results, XGBoost demonstrates the smallest MSE (1.377665), signifying it generates the most precise predictions, while Random Forest shows the greatest error (3.608123).

MAE indicates the average absolute deviation between real and forecasted values, offering a clear understanding of prediction errors.

MAE is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (2)$$

Once more, XGBoost records the smallest MAE (0.85685), validating its capability to reduce prediction errors more effectively than Linear Regression (1.041979) and Random Forest (1.517053).

RMSE is the square root of the mean squared error and is utilised to assess the size of prediction errors in the same unit as the target variable.

It is conveyed as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (3)$$

RMSE increases the impact of larger errors because of the squaring, rendering it more responsive to substantial discrepancies. XGBoost exhibits the lowest RMSE (1.17374), whereas Random Forest shows the highest (1.899506), suggesting that XGBoost produces the least significant prediction mistakes.

R^2 quantify the extent of variance in the target variable that the model accounts for. The formula is given as:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \qquad (4)$$

A higher R^2 value (closer to 1) indicates better model performance. XGBoost achieves the highest R^2 score (0.810543), demonstrating superior explanatory power, while Random Forest has the lowest (0.503809), showing weaker predictive capability.

According to these assessment metrics, XGBoost clearly surpasses both Linear Regression and Random Forest in every respect. It attains the minimum MSE, MAE, and RMSE, simultaneously achieving the highest R^2 score, signalling greater predictive precision and reliability. Based on these findings, XGBoost has been chosen as the ultimate machine learning technique for this research.

Association Rules Mining

To analyse consumer buying habits and produce relevant product suggestions, the recommendation system was developed utilising Association Rule Mining, specifically the Apriori algorithm. The Apriori algorithm identifies which items are commonly purchased together by finding frequent itemsets in transaction data and generating association rules. This strategy assists businesses in improving their cross-selling and up-selling tactics by suggesting complementary products. It is especially useful for market basket analysis.

Figure 5 enhances the recommendation system by offering various alternative choices for each item. Frequent itemset mining is carried out, and support factors are calculated; the relevancy of recommendation items is determined by confidence. There are multiple recommended items for dealers to choose to promote to their customers thus increasing their success in upselling hit rates. After determining frequent itemsets using the confidence factor, association rule mining is completed, supported the relevancy of recommended items. This system enhances the dealers' marketing strategies and decisions by offering data-driven insights.

Power BI Dashboard

This Power BI solution contains a complete data analysis channel. The objectives are business improvement, product recommendations and sales forecasts. The two key elements in this dashboard are outcomes from predictive models and recommendation system rules. These outcomes are the input for this solution, KPI are derived from these data. This Power BI dashboard provides default and customised analysis controlled by users through the interactive interface.

	Item	Recommendation 1	Recommendation 2	Recommendation 3	Recommendation 4
0	AUTOMOTIVE	HOME AND KITCHEN II	EGGS	FROZEN FOODS	LADIESWEAR
38	BABY CARE	LAWN AND GARDEN	BREAD/BAKERY	PREPARED FOODS	LADIESWEAR
76	BEAUTY	PERSONAL CARE	HOME CARE	HOME APPLIANCES	PRODUCE
114	BEER	LIQUOR, WINE	WINE	LIQUOR	PRODUCE, LIQUOR
245	BEVERAGES	BEER, WINE, LIQUOR	LIQUOR, WINE	BEER	BEER, WINE
283	BOOKS	LADIESWEAR	BREAD/BAKERY	GROCERY II	DAIRY
321	BREAD/BAKERY	PREPARED FOODS	LAWN AND GARDEN	SCHOOL AND OFFICE SUPPLIES	POULTRY
359	CELEBRATION	MEATS	DAIRY	PET SUPPLIES	WINE
397	CLEANING	SCHOOL AND OFFICE SUPPLIES	MAGAZINES	BEAUTY	LAWN AND GARDEN
435	DAIRY	BREAD/BAKERY	HOME AND KITCHEN II	SCHOOL AND OFFICE SUPPLIES	BOOKS
473	DELI	MEATS	CELEBRATION	PERSONAL CARE	PLAYERS AND ELECTRONICS

Figure 5 Frequent itemset mining

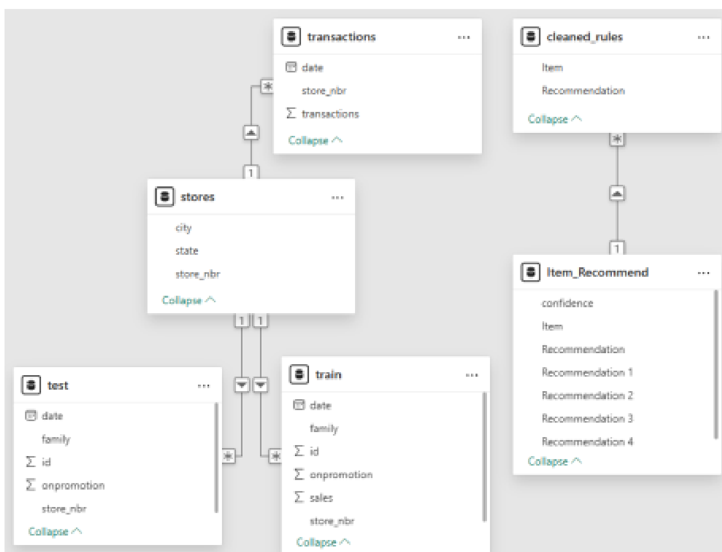

Figure 6 Power BI model view

Figure 6 shows the Power BI model view, the links between tables represent their relationship and connections with each table. The core variable that connects the table group at left, which is the predictive model output, is store id, which is store_nbr. The core variable that connects the two more tables at the right is Item, which represents the product chosen by dealers when operating the Power BI dashboard.

Figure 7 shows the Power BI dashboard interface. The Power BI dashboard offers a detailed summary of sales outcomes, suggested products, and regional sales trends. Highlighted KPIs include total revenue (1.07 billion), the most successful store location (store 44), and the leading product categories by sales (e.g., Dairy, Cleaning, Beverages). The dashboard's dynamic suggestion system improves opportunities for cross-selling and up-selling by enabling customers to choose a product category and see recommended related items such as "Personal Care," "Home Care," "Home Appliances," and "Produce."

The dashboard features filters for regions, cities, and locations, allowing for comprehensive analysis across multiple levels. A time-series analysis monitors sales trends between 2013 and 2017, while a geospatial visualisation illustrates store locations and sales distributions across cities. These visual insights can assist business executives in enhancing client engagement (tracking customer journey data), sales strategies

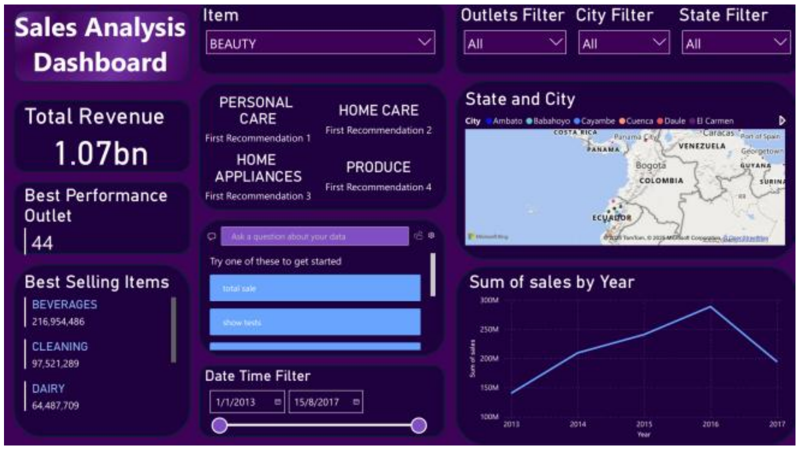

Figure 7 Power BI interface

(performance by region or country), and revenue growth (data visuals about product demand).

This research provided evidence to prove that by integrating predictive analytics and recommendation system association rules, this Power BI dashboard serves as a one-stop data-driven solution, namely a Next-Gen DMS. This DMS provides actionable insights to stakeholders such as management teams, executives, dealers, and distributors. The dashboard provides convenience by catering to different users, thus assisting users in upselling and cross-selling activities.

CONCLUSION

This research addressed the objectives by providing suitable action plans and execution methodologies to create a predictive data analytics model to promote upselling and cross-selling activities based on consumer buying patterns. A Next-Gen Dealer Management System (DMS) has been successfully developed through this research. By investigating and evaluating different machine learning and data mining models, the best-performing models identified are Extreme Gradient Boosting (XGBoost) for predictive models and Association Rule Mining (Apriori algorithm) for recommendation systems. Through the execution methodologies defined in this research, the predictive model and recommendation systems are developed and integrated into a dynamic and interactive Power BI dashboard. The three predictive models developed are evaluated using performance metrics, with XGBoost being the most suitable predictive method for this research. The dashboard provides dealers with actionable insights and relevant product recommendations to aid the upselling and cross-selling activities and increase overall sales performance. This research provides a centralised DMS to help businesses stay competitive in this fast-paced market. The DMS solution developed in this research is proven to be scalable using three different synthetic datasets, while XGBoost is the leading machine learning model. This research can serve as a baseline solution for new researchers to encapsulate more research aspects, such as integrating AI Agents, workflow automations, machine learning optimisers, and exploring more data analytics approaches. Industrial data analytics should be considered to ensure the solution delivered is industrial-ready and fulfils industrial needs.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Universiti Teknologi PETRONAS and Department of Computer and Information Science for providing the facilities and support required for this research. The authors also thank the anonymous reviewers for their constructive feedback, which greatly improved the quality of this article.

REFERENCES

[1] S.E. Schaeffer and S.V.R. Sanchez, "Forecasting client retention — A machine-learning approach," *J. Retail. Consum. Serv.,* vol. 52, p. 101918, 2020.

[2] O. Dogan, "A recommendation system in e-commerce with profit-support fuzzy association rule mining (p-farm)," *J. Theor. Appl. Electron. Commer. Res.,* vol. 18, no. 2, pp. 831-847, 2023.

[3] S. Lee, S. Ko, A.H. Roudsari, and W. Lee, "A deep learning model for predicting the number of stores and average sales in commercial district," *Data Knowl. Eng.,* vol. 150, p. 102277, 2024.

[4] J. Herce-Zelaya, C. Porcel, J. Bernabé-Moreno, A. Tejeda-Lorente, and E. Herrera-Viedma, "New technique to alleviate the cold start problem in recommender systems using information from social media and random decision forests," *Inf. Sci.,* vol. 536, pp. 156-170, 2020.

[5] M. Mustak, H. Hallikainen, T. Laukkanen, L. Plé, L. D. Hollebeek, and M. Aleem, "Using machine learning to develop customer insights from user-generated content," *J. Retail. Consum. Serv.,* vol. 81, p. 104034, 2024.

[6] A.M. Pereira, J.A.B. Moura, E.D.B. Costa, T. Vieira, A.R.D.B. Landim, E. Bazaki, and V. Wanick, "Customer models for artificial intelligence-based decision support in fashion online retail supply chains," *Decis. Support Syst.*, vol. 158, p. 113795, 2022.

[7] B. Walek and P. Fajmon, "A hybrid recommender system for an online store using a fuzzy expert system," *Expert Syst. Appl.*, vol. 212, p. 118565, 2023.

[8] H. Paredes-Frigolett and L.F.A.M. Gomes, "A novel method for rule extraction in a knowledge-based innovation tutoring system," *Knowl.-Based Syst.,* vol. 92, pp. 183-199, 2016.

[9] R.H. Lin, W.W. Chuang, C.L. Chuang, and W.S. Chang, "Applied big data analysis to build customer product recommendation model," *Sustainability*, vol. 13, p. 4985, 2021.

[10] J. Gasper and J. Cahalan, "Utilising the random forest algorithm and interpretable machine learning to inform post-stratification of commercial fisheries data," *Fish. Res.*, vol. 281, p. 107253, 2025.

[11] G. Feng and M. Fan, "Research on learning behavior patterns from the perspective of educational data mining: Evaluation, prediction and visualisation," *Expert Syst. Appl.,* vol. 237, p. 121555, 2024.