Utilizing Predictive Analytics to Enhance Retail Business Performance

Aravind Hallimysore Kalegowda
MSc in Data Analytics
X22104275
National College of Ireland

Viva Question and Answers

1. In threshold-based approach that was used in data cleaning. What exactly it is and how it was used in data a cleaning step?

Ans: The threshold-based approach in data cleaning is setting specific limits to identify and eliminate outliers from the data. In this study this approach was applied to key variables like 'Quantity' and 'Unit Price' in the dataset. By setting thresholds, values that were exceptionally high or low (which could indicate errors or atypical cases) were identified and removed. This helped in creating a more accurate and representative dataset for analysis and this method is particularly useful in retail data. This approach was particularly beneficial in enhancing the accuracy and representativeness of the dataset for subsequent analysis with a specific emphasis on its applicability to retail data. Particularly the thresholds were determined using the 1st and 99th percentiles providing a robust method for deciding which values to drop. Using percentiles in a systematic manner ensured a thorough and data-driven strategy for dealing with extreme values by enhancing the overall dependability of the dataset.

2. “CustomerID” column has a substantial 24.93% missing values. How were they handled?

Ans: Missing values in the “CustomerID” column were dropped because of the following reason. Despite of dropping 24.93% rows still the number of rows in dataset was “406829” which sufficient to conduct the experiment. Hence 24.93% of data was removed.

Dropping rows with missing "CustomerID" values to maintain data accuracy and relevance in customer segmentation analysis. Keeping rows without a customer ID could affect data integrity and lead to misleading conclusions. Especially in the precise identification of individual behaviours for segmentation. Imputing unique customer IDs was not practical, so in this study it has opted for the simplicity of dropping such rows instead of using more complex methods. This decision focuses on ensuring accuracy and clarity in the dataset promoting a transparent data cleaning process and boosting the reliability of subsequent customer segmentation analyses.

3. What is the difference between prediction and estimation? Elaborate with reference to your objective of “store sales prediction”.

Ans:

Prediction: Prediction involves forecasting future outcomes based on existing data. It's about using patterns in the data to project these patterns into the future. In the "store sales prediction" study using historical sales data and other influencing factors to anticipate future sales, such as predicting the volume for the upcoming month. The machine learning models in the "Store Sales Prediction" study illustrate this by generating sales forecasts based on past data and relevant features and also those models accuracy is evaluated by using RMSE, R^2 score, MSE and MAE metrics.

Estimation: Estimation is about inferring the values of unknown parameters in data model. It’s about understanding the relationships between different variables in data. In the "Store Sales Prediction" study estimation would involve determining how various factors like store area, Items available, Daily customer count and store sales influence over all sales.

Prediction anticipates future outcomes, while estimation delves into the understanding of underlying data relationships and parameters both playing essential roles in informed decision-making in the store sale prediction study.

4. What is the difference between Random Forest technique and Decision Trees? Why is there such a large difference (nearly 6 times in RMSE values) between the metrics reported for Random Forest regressor and Decision Tree regressor, with Random Forests performing significantly worse?

Ans:

The main difference between Random Forest and Decision Trees lies in their structures and functions. Decision Trees are straightforward models that split data based on feature values, while Random Forests are ensembles of Decision Trees designed to enhance prediction accuracy. Random Forests address overfitting issues associated with Decision Trees by averaging multiple trees providing a better bias variance adjustment.

Model Name	R-Squared Value	MSE	RMSE	MAE
XGB Regressor	0.97	8651550.52	2941.35	2181.04
LGBM Regressor	0.97	6867481.82	2620.59	2113.72
Cat Boost Regressor	**0.98**	7485044.54	2735.88	2116.07
Random Forest Regressor	**0.98**	**6283989.01**	**2506.79**	**2001.23**
Decision Tree Regressor	0.96	11514901.69	3393.36	2721.52
Linear Regression	0.00	270688558.55	16452.62	13125.63

Fig.1. **Comparing outcomes of Machine Learning Models**

In this research The RMSE value of Random Forest is better than Decision tree (Fig.1.) and there no such big difference (6times) in RMSE value between random forest and Decision tree. The significant difference in RMSE values between the Random Forest regressor and Decision Tree regressor is mainly due to the nature of their algorithms. Decision Trees being single trees can struggle with complex data and are prone to overfitting, learning noise along with patterns. Where in Random Forest regressor which combines predictions from multiple trees is more robust especially in handling overfitting and complex relationships in the data. The RMSE difference indicates that for this specific project is the ensemble approach of Random Forest is more effective in capturing the dataset's complexities, leading to better predictions compared to a single Decision Tree.

5. How did you use k-means clustering for store sales prediction (Section 6.3.2)? What are the metrics for this technique?

Ans: K-means clustering in sale prediction was basically used for preliminary analysis and understanding the data, this section details the process of applying k-means clustering to g roup store data based on various attributes such as item ratios, store size, and combination s of these factors. The approach aimed to form clusters that provide insights into customer segments and support store management strategies. The Elbow Method was utilized to det ermine the optimal number of clusters and used in creating new variables such as “Labels ” and “Target Groups”. But in this study k-means clustering has not used further executio n. Hence no specific metrics for k-means clustering was used in this store sales prediction. “Silhouette score” was used as evaluation metrics to measure the K-means clustering in th e “Customer segmentation” research.