OMNIe Solutions – Task 02 Project Evaluation Report

Name: Dipin Raj
University: Chandigarh University
Ph. No: 6235876977
E-mail: dipinr505@gmail.com
Wayand, India- 673593

Assignment: Evaluation 01
Job Role: AI/ML Intern
Submitted To: Mr. Chhatra Pratap & Mr. Anand Shukla
Company Name: Omnie Solutions
Noida, India

I. TASK 1: PROBLEM STATEMENT

Dataset: Retail Web Session Intelligence (RWSI)

The Retail Web Session Intelligence (RWSI) dataset simulates customer interactions on a digital retail platform that sells consumer and lifestyle products.

- Each record represents an anonymized user session, capturing browsing patterns, engagement metrics, contextual attributes, and conversion outcomes.
- The goal is to build predictive and diagnostic models that help understand what drives successful purchase intent and user engagement.

Problem Statement: Develop a model that can predict the likelihood of a conversion event (MonetaryConversion) based on a user's browsing behavior, engagement metrics, and contextual factors.

II. TASK 1: DELIVERABLES

To Understand digital behavior data — explore session-level features such as browsing patterns, engagement metrics, and contextual variables (e.g., region, month, traffic source).

- Perform exploratory data analysis (EDA) — identify trends, correlations, and anomalies; visualize how behavior differs between converting and non-converting sessions.
- Handle missing values
- Engineer new features
- Build predictive models
- Evaluate performance
- Interpret model results
- Communicate findings — summarize insights into what differentiates high-intent shoppers from casual browsers.

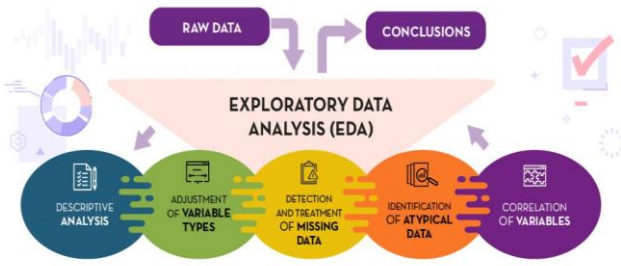

Fig 1: EDA (Exploratory Data Analysis)

II. DATASET ANALYSIS

The data is composed of logs of user interactions and conversion on a marketing/sales platform. The individual sessions are represented in each of the rows and consist of numerical measures of engagement and categorical features of the user and his or her environment.

- Numerical aspects: 'pageEngagementScore', 'visitDuration', 'click rate', 'session Count', etc.
- Categorical/contextual: 'visitMonth', 'userCategory', 'marketZone', region, and conversion target (Yes/No).
- Metadata: 'sessionID', 'timestamp', among others.

The dataset had a low percentage of missing values that though small as compared to the overall amount of data was well managed to avoid data leakage. Another important point that had to be made was the high-class imbalance (around 85:15) between non-converting and converting users that forced to employ class-balancing strategies when preprocessing.

III. EXPLORATORY DATA ANALYSIS (EDA)

The EDA step has brought out some of the major behavioral and structural trends in the data. There was a clear imbalance in the target variable as the positive conversion instances were significantly fewer than the negatives. The numerical variables like 'pageEngagementScore' and 'timeOnPage' were skewed to the right meaning that a few users had a much deeper engagement than the mean.

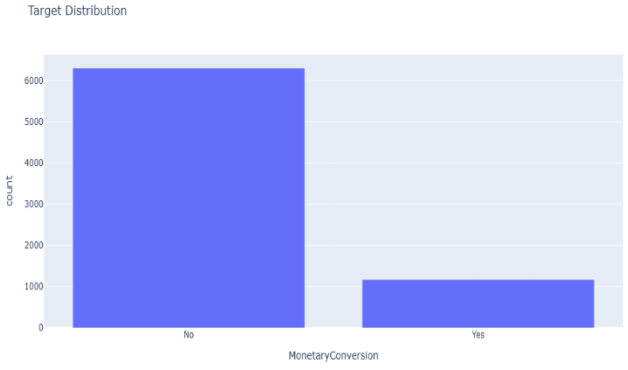

Fig 2: Graph revealing class imbalance

There were strong correlations between 'pageEngagementScore' and conversion variable indicating that user conversion likelihood is strongly correlated with higher engagement. Two of the categorical variables, 'visitMonth' and 'userCategory' had different conversion patterns - conversions tended to rise at specific months and to be more among a specific user segment which is expected of marketing cycles.

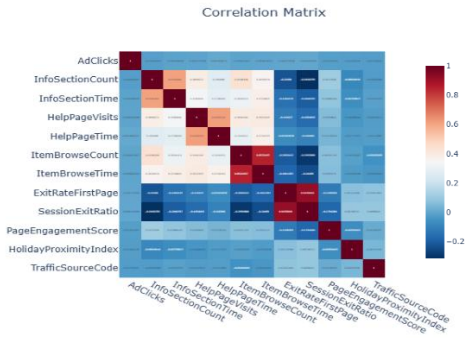

Fig 3: Correlation among independent variables

There were outliers in engagement-related metrics, however, most of them seemed to be the actual high-activity users and not mistakes. On the whole, the information demonstrated that the intensity and timing of user interactions was the most dominant factor affecting conversions.

IV. METHODOLOGY

Data Preprocessing was done to narrow down data quality and maintain the underlying patterns of behavior. Models were pruned of null values in order to guarantee integrity and consistency of model inputs. Outlier management was done using a graphical, visual, instead of an automated, manner. Plotting feature specific limits was done manually using a scatter plot - in which the true extreme behavior would be retained whilst all the points that distorted the distribution symmetric would be discarded. This did not sacrifice significant variance in the stability of the model.

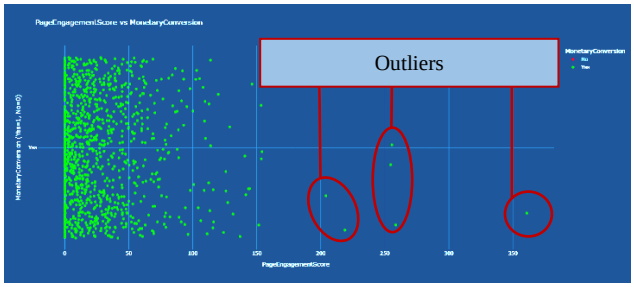

Fig 4: Defining outlier mitigation range

In order to counter the imbalance of classes, a range-based semi-synthetic over-sampling policy was adopted. New Yes cases were created as a result of controlled randomness rather than direct duplication within the previously determined valid range of key features. It was used to make the positive classes more diverse, enhance model generalization, and avoid the formation of bias.

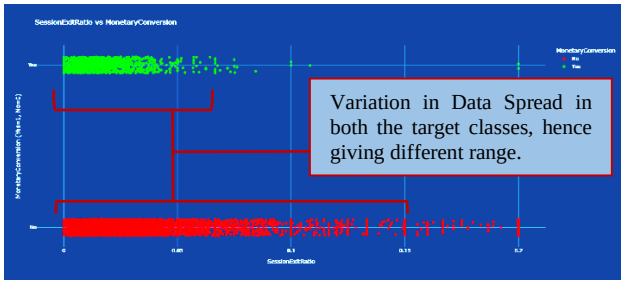

Fig 1: Class-Wise Variation in Data Spread and Outlier Boundaries

Numerical compatibility and one-hot encoding of categorical features was implemented, and feature scaling was used to ensure similarity among the magnitudes of variables. Post-encoding correlation checks confirmed previous observations and ‘pageEngagementScore’ and ‘visitMonth’ are still indicated to have a dominant predictive power. These preprocessing processes were systematic and were done in a methodical manner and put together resulted in a stable well-balanced and behavior-preserving dataset, which served as a robust foundation to base on model development.

V. RESULTS

The experimental analysis showed the mixed degree of the performance of the implemented models with Random Forest and XGBoost models becoming the most predictable ones regarding conversion results.

The consistency of training stability was ensured by the fact that loss curves were always smooth and that the recall and F1 curves were parallel to each other on both classes. Nonetheless, the ROC curves showed comparatively slow convergence which implied that the dataset did not have very distinctive patterns required to make very certain classification.

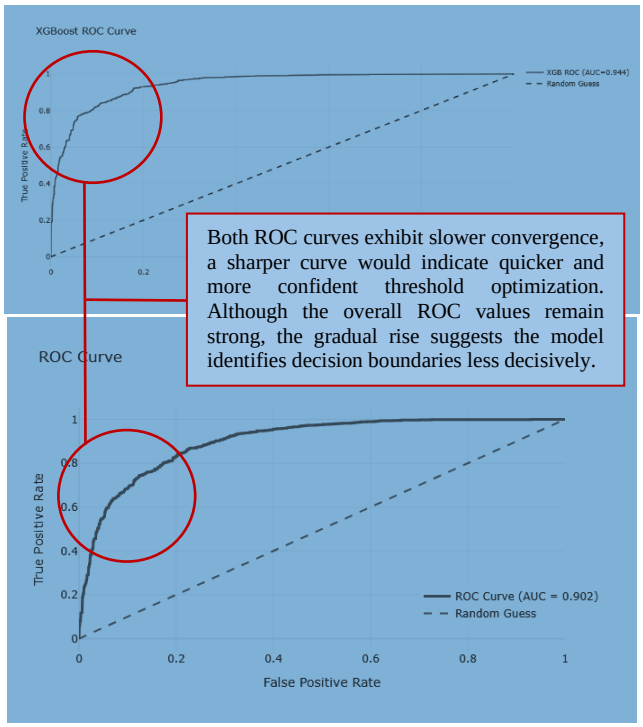

Fig 1: ROC curve showcasing slower convergence

It means that although the models were able to model the general behavioral tendencies, some hidden or composite dependencies might still be underrepresented by the existing set of features.

The model that had the most stable and balanced results was the Random Forest with the best balance between precision and recall. This was due to its strong ability to control both types of classes, which can be attributed to the positive effect of the semi-synthetic balancing technique, which was efficient in reducing the bias towards the majority class and increasing the generalization ability of the model.

Table 1: Model Performance Summary

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
XGBoost	0.865	0.87	0.86	0.86	0.94
Random Forest	0.867	0.87	0.87	0.87	0.94
Decision Tree	0.865	0.87	0.87	0.87	0.94
Logistic Regression (Scratch)	0.810	0.81	0.81	0.81	0.90
Logistic Regression (Library)	0.815	0.82	0.82	0.81	0.91

VI. CONCLUSION

This research was able to prove that it is possible to find statistical correlation between prediction of conversion and behavioral and engagement-based features. As it was pointed out in the analysis, the predictive power of the variables like 'pageEngagementScore' and 'visitMonth' is the largest one, as in real-world marketing, it is the real-time and the intensity of interaction that leads to the conversion.

A semi-synthetic balancing methodology and range-specific outlier process turned out to be useful to maximize robustness of the model and not to distort the natural data behavior. The above preprocessing strategies made the learning process consistent and cross-logistically interpretable.

Random Forest and XGBoost were the most successful models of the tested ones in terms of accuracy and generalization. The feature influence was confirmed by the Logistic Regression, although it was less accurate with the advantage of interpretability.

Altogether, as noted in the project, data quality, careful preprocessing and feature intuition are as important as the complexity of the model. As datasets get richer and more time can be tracked, this framework can transform into a real-time conversion prediction system helping to execute strategic marketing and optimizing resources.

REFERENCE

[1] Colab Notebook Link: Retail Web Session Intelligence (RWSI)

[2] GitHub Repo Link: Predicting Operational Efficiency of Manufacturing Teams

[3] Cousineau, D. and Chartier, S., 2010. Outliers detection and treatment: a review. *International journal of psychological research*, *3*(1), pp.58-67.

[4] Iduseri, A., 2022. Winsorization with Graphical Diagnostic for Obtaining a Statistically Optimal Classification. *Advances in Principal Component Analysis*, p.199.