Big Data Analysis and Project

Assignment 1: Part D (Report)

Possakorn a1873765

Part 1: Restatement and Summary

Over recent years, the world of eCommerce platforms has expanded and changed, largely because of new technology and the Coronavirus disease situation which change the way people now prefer to shop. When we look at the huge amount of information from their online shopping platforms, like purchase details and customer feedback, we see it offers a great opportunity for businesses to improve and increase their profits (Vanaja and Belwal, 2018). One of the important areas is understanding how customers feel, as this can give clues about potential earnings. A detailed study using data from a well-known Brazilian online store tried to find out which website features made customers leave (Olist and Sionek, 2018). LightGBM method was the most successful algorithm in this analysis which provide some main reasons customers left related to their last purchase duration, spending behavior, location, and the scores they gave in reviews.

Restatement the question following the list below;
01 How does sentiment impact business profitability in the E-commerce industry for each segment?
02 How does sentiment in customer feedback affect the customer lifetime value in the E-commerce sector, especially within specific customer segments?
03 What are the key platform features that most effectively minimize customer churn?

Part 2: Analysis and Visualisation

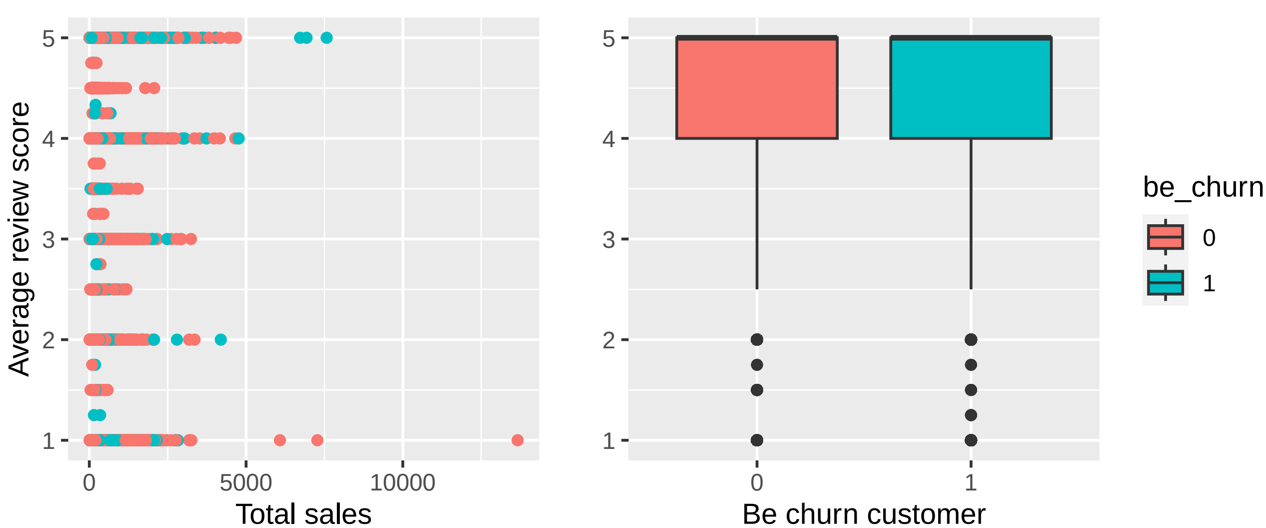

Figure 1: Relationship between customer life time value and average review score

To derive the desired outcomes, a structured methodology anchored in foundation data science principles was implemented. At the outset, the data underwent an intensive preparation phase, where several data sets were intricately combined. This consolidation facilitated a preliminary exploratory data analysis, enabling us to discern any immediately evident patterns or trends. Interestingly, a closer examination of **Figure01** revealed that review scores seemed to be relatively consistent, showing minimal variation, especially in the case of customers who ceased their engagements or 'churned'.

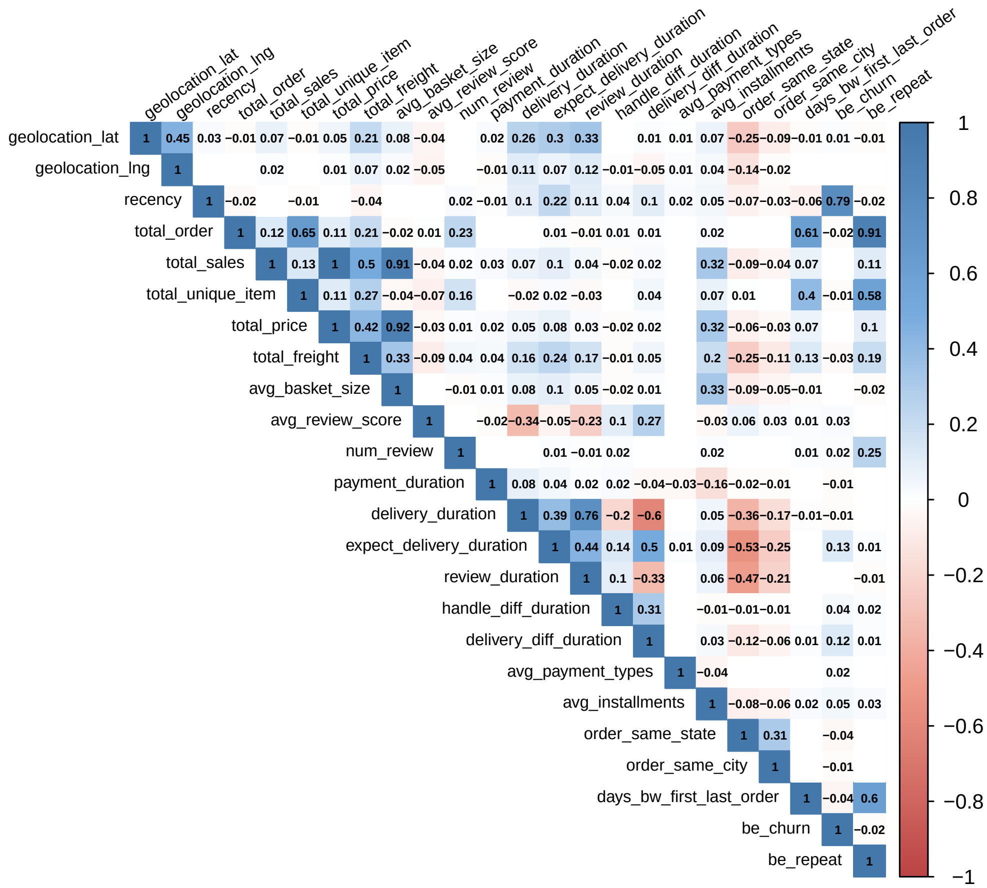

Figure 2: correlation matrix represent the relattionship between platform feature and customer churn

Recognizing the need for a deeper understanding of how various data variables interacted, a comprehensive correlation plot was constructed. This visual representation, depicted in **Figure02**, offers a holistic view of the intricate interrelationships among the diverse dataset variables. Armed with insights from this correlation analysis, we meticulously selected certain data features. Based on this analysis, specific features were then chosen to be tested using several machine learning models like logistic regression(Jain, Khunteta and Srivastava, 2020), Support Vector Machine(Y., 2022), Light Gradient Boosting Machine(Ke *et al.*, 2017), and eXtreme Gradient Boosting(Chen and Guestrin, 2016).

Table 1: Model performance comparing before and after tuning

	Before Tuning			After Tuning		
Model	Accuracy - train	Accuracy - test	ROC AUC	Accuracy - train	Accuracy - test	ROC AUC
lgbm	1.000	0.994	0.477	0.989	0.989	0.502
logistic	0.993	0.994	0.399	0.993	0.994	0.398
svm	0.993	0.994	0.494	0.993	0.994	0.427
xgboost	0.993	0.994	0.430	0.993	0.994	0.447

Upon reviewing the outputs, as tabulated in **Table01**, it became evident that the LightGBM model outperformed its counterparts. Not only was it demonstrably faster, but it also showcased commendable adaptability with extensive datasets. In the final analytical phase, we harnessed this model to pinpoint the most impact features, the specifics of which are elaborated upon in **Table02**.

Table 2: Individual feature importance on each attribute (top features).

Variable	Importance
recency	0.301
net_sales	0.196
payment_duration	0.194
average_installments	0.175
state_SP	0.080
avg_review_score	0.031
order_same_state	0.024

Part 3: Improvement of Situation

In the analysis outlined in **Table01**, a multifaceted approach was adopted to ensure the accuracy and integrity of the model's results. A critical initial step involved feature engineering and feature selection. These methodologies allow for the refinement of the dataset, ensuring that the most relevant variables are prioritized and selected for input into our computational model.Subsequent to this, the following dataset performs a rigorous preprocessing which encompassed a series of operations including normalization (to scale data values), encoding (to convert categorical data), the elimination of zero-variance (to remove non-informative variables), and addressing collinearity (to prevent redundant features from skewing results).

Table 3: Hyperparameters and ranges used during grid search

Classifier	Hyperparameter	Values	tuned_values
Logistic Regression	Regularization	[L1, L2 regularization]	L2
	Penalty term	tune()	1e-10
Suport Vector Machine	Kernal	Radial Basis Function(RBF)	
	Cost	tune()	2.378
	Rbf_sigma	tune()	1e-05
LightGBM	Trees	default = 1000	
	mtry	range[1, floor(sqrt(number of feature))]	1
	min samples per leaf	range[1, 10]	5
	learning rate	range[0.01, 0.3]	1.023
XGboost	Trees	default = 1000	
	mtry	range[1, floor(sqrt(number of feature))]	3
	min samples per leaf	range[1, 10]	1
	learning rate	range[0.01, 0.3]	1.023

A notable challenge was the pronounced imbalance within the dataset. Rather than employing undersampling or oversampling techniques, which often risk introducing bias or overfitting, a decision was made to maintain the dataset's original state. Given the temporal nature of the data, a 'time slicing' strategy was employed (Gattermann-Itschert and Thonemann, 2021). This method partitions data into distinct chronological segments for training and validation, ensuring temporal integrity and preventing potential data leakage. Lastly, the refinement of the model was done via hyperparameter tuning. This was achieved using the k-fold cross-validation technique, a practical standard in enhancing model efficiency and controlling the behavior of the machine learning algorithm, as illustrated in **Table03**.

Part 4: Conclusion and Future Work

Following the empirical performance from the preceding analyses, the LGBM model distinctly stands out. Its supremacy is not merely confined to its excellent speed; the model showcases robust adaptability even when grappling with large and uneven datasets. However, a limitation becomes apparent when examining its performance. The model appears somewhat constrained, primarily due to data paucity on specific customer behaviors, especially those who engage in singular transactions. This limitation is accentuated by the inherent challenges in sourcing extensive data, given its private and specific nature.
As we plan our next steps in research, it's essential to look at the broader perspective. While customer-centric approaches offer valuable insights for revenue optimization, the seller's perspective remains an crucial aspect of the e-commerce ecosystem. In online shopping, Sellers often display a more consistent transaction behavior, providing a stable foundation for analysis. By understanding the seller's side, we might discover new insights that can help keep users active and grow the online platform.

(word count: 794 - exclude Table, Reference, and Appendix)

Reference

Chen, T. and Guestrin, C. (2016) "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* New York, NY, USA: Association for Computing Machinery (KDD '16), pp. 785–794. Available at: https://doi.org/10.1145/2939672.2939785.

Gattermann-Itschert, T. and Thonemann, U.W. (2021) "How training on multiple time slices improves performance in churn prediction," *European Journal of Operational Research*, 295(2), pp. 664–674. Available at: https://doi.org/https://doi.org/10.1016/j.ejor.2021.05.035.

Jain, H., Khunteta, A. and Srivastava, S. (2020) "Churn prediction in telecommunication using logistic regression and logit boost," *Procedia Computer Science*, 167, pp. 101–112. Available at: https://doi.org/https://doi.org/10.1016/j.procs.2020.03.187.

Ke, G. *et al.* (2017) "LightGBM: A highly efficient gradient boosting decision tree," in I. Guyon et al. (eds.) *Advances in neural information processing systems.* Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Olist and Sionek, A. (2018) "Brazilian e-commerce public dataset by olist." Kaggle. Available at: https://doi.org/10.34740/KAGGLE/DSV/195341.

Vanaja, S. and Belwal, M. (2018) "Aspect-level sentiment analysis on e-commerce data," in *2018 international conference on inventive research in computing applications (ICIRCA)*, pp. 1275–1279. Available at: https://doi.org/10.1109/ICIRCA.2018.8597286.

Y., T.V.A.S., Nguyen Nhu AND Ly (2022) "Churn prediction in telecommunication industry using kernel support vector machines," *PLOS ONE*, 17(5), pp. 1–18. Available at: https://doi.org/10.1371/journal.pone.0267935.