CHAPTER 5

DISCUSSION AND FUTURE WORK

5.1 Introduction

This chapter discusses the results generated from customer churn prediction in e-commerce industry. This chapter begins with dataset identification, followed by Exploratory Data Analysis (EDA). Then, data preprocessing and feature engineering is comprehensively approached. Logistic Regression, Random Forest, Random Forest attached with SMOTE, and XGBoost are used to evaluate the best model. After that, hyperparameter tuning is done on all 4 models. The purpose is to investigate the changes between 4 models, before and after hyperparameter tuning. According to the results generated during model implementation, it is proven that hyperparameter tuning would improve model performance across multiple metrics. The improved results would be effective in developing robust customer churn prediction models for e-commerce applications. The aim of making a positive contribution to customer retention strategies in e-commerce industry could be guaranteed.

5.2 Summary

Customer churn prediction in e-commerce industry aims to identify customers who are likely to stop using the platform services. This analysis clarifies customer behavioral patterns through data taken from an e-commerce dataset containing 5,630 customer records. This project involves several phases, from data collection to final model evaluation and business impact analysis.

The dataset initially contained class imbalance where 83.2% were non-churned customers and 16.8% were churned customers. The data after collection went through a cleaning stage which means that various preprocessing techniques were carried out, including

handling missing values, removing duplicates, and feature engineering. New derived features were created such as Engagement Score, Customer Value Score, and Recency Score to better capture customer behavior patterns. Then, categorical encoding and numerical scaling were performed to create uniform data ready for machine learning algorithms.

The processed data was then used to train four different machine learning algorithms: Logistic Regression, Random Forest, Random Forest with SMOTE, XGBoost. Out of all models before hyperparameter tuning, it is shown that XGBoost performs the best, which generates an F1-Score of 0.8456. Out of all models after hyperparameter tuning, it is shown that Random Forest performs the best, which generates an F1-Score of 0.8556. Out of all 4 models, only Random Forest, Random Forest with SMOTE, and XGBoost achieves the minimum benchmark of F1-Score, which is 77%. Unlike Logistic Regression, it does not achieve the minimum benchmark because the F1-Score gained before and after hyperparameter tuning is 0.5952 respectively. This demonstrates that Logistic Regression is not the suitable machine learning algorithm for customer churn prediction.

With a deeper analysis, it can be seen that behavioral factors such as complaints, tenure, and satisfaction scores are more important predictors than demographic information. Feature correlation analysis revealed that tenure shows strong negative correlation (-0.338) with churn probability, reflecting that longer-tenure customers are less likely to churn.

From the success of the project, we can draw the following conclusions. The first conclusion is that data quality and feature engineering are crucial. Proper data preprocessing and derived feature creation provides better results compared to model training without data preprocessing. The second conclusion is that XGBoost and Random Forest performs the best before and after hyperparameter tuning respectively. The third conclusion is that churn prediction system has high potential in improving the ROI by utilizing customer retention strategies.

Overall, the project successfully achieved its goal of developing effective customer churn prediction models in a structured and data-driven manner. The project also demonstrated that machine learning-based churn prediction can act as a powerful tool in supporting customer retention decisions in real-time business operations.

5.3 Future Works

The customer churn prediction analysis revealed critical patterns and insights from the comprehensive examination of 5,630 customer records. The systematic analysis identified key behavioral indicators and established the foundation for predictive model development through statistical examination of churn relationships. The suggested future works would be listed down as shown below.

a) Larger dataset volume

   The current dataset only contains 5630 rows of data, which is insufficient in model training. In the production scenario, 5630 rows of data are considered little, because the dataset in production environment starts from millions. It indicates that more meaningful patterns could be extracted out when the information is sufficiently enough.

b) Application of Deep Learning Models

   This case study applies traditional machine learning models. But traditional machine learning models have its limit in approaching to more complex relationships. To make a breakthrough on this barrier, Deep Learning models would be a better solution, because Deep Learning models are designed to find out the complex relationship from the dataset.

c) Real Time Implementation and Monitoring

   Current solution does not include production deployment. This indicates that the result gained from the training does not exactly reflect the real-time scenario. The differences between local testing and production execution could be significantly different. When the model does not reveal the mistakes, it is hard to find out the limitations in the current solution. Only the production environment would greatly reveal the flaws in the current solution.

d) Enhanced Feature Engineering with external data sources.

   The solution only use 1 type of dataset. The features extracted from single source of dataset would be insufficient to make insightful decisions. Additional dataset

such as transactional data from e-commerce industry should be used, as the transaction activity represents the real-time action, which proves the record is valid.

The above steps will allow further research to increase the scope of this project, improving the accuracy and practical relevance of customer churn prediction systems. The current project has paved the way for the use of machine learning as an effective customer retention tool in e-commerce; thus, further development will have greater implications in the future for strategic business decision-making and customer relationship management.