Light Gradient Boosting Machine for E-commerce Customer Churn Prediction

by Vanessa Atta-Fynn - 10665130

INTRODUCTION

- **Background & Motivation**

Customer churn or customer attrition is what happens when the relationship of a customer with a company or service comes to an end. Customer churn is a phenomenon that affects all kinds of industries and sectors of work from telecommunication, retail and e-commerce, finance to airline markets.

Customer churn prediction is one of the many methods developed to improve customer retention. It predicts the likelihood of customer churn and improves retention by honing in on likely abandoners. With this knowledge businesses are given a chance to take remedial actions to reduce or prevent churning.

Though e-commerce and retail services records one of the highest churn rates as shown in Figure 1, not much attention is paid to customer churn in e-commerce websites [1].

- **Research Objective**

This study seeks to build and test algorithms for predicting customer churn in e-commerce sector by using a sample data set within this field.

- **Contribution of Study**

At the time of publication, there was a noticeable research gap in the analysis and prediction of customer churn in the retail and e-commerce sector in Ghana. Most retail business owners and e-commerce stakeholders focus more on marketing strategies but are quite oblivious to this innovation. This paper serves as a stepping stone to breach that gap. This study is also geared towards contributing to the retail and e-commerce sector of the economy of the country by creating an alternative model to predict customer churn, thereby giving end-users the chance of making active decisions and remedial actions.

PROBLEM STATEMENT

Customer churn is very common but also very expensive for businesses.

The lack of understanding of customers and their needs makes it difficult for them to maintain customer loyalty. Due to the non-contractual relationship in e-commerce, it is relatively difficult to predict customer churn and reasons behind it through traditional statistical analysis.

- **Limitation of Study**

This study focuses primarily on the e-commerce sector of the retail industry, as it is rather tedious to apply it to the offline world as this study is based on customer behavior on the web.

METHODOLOGY

- **Dataset Description**

As inspired by [2] a data set with adequate customer details was collected for this study. The data set consist of a range of customer data collected as well as a churn flag showing which customers churned and which customers did not churn. It consists of a total of 20 columns and 5630 rows.

- **Data Preprocessing**

First of, the problem of missing values in column was identified and handled. Next, the data set was observed to have some irrelevant columns that would not play any role in bringing insights, hence these columns were drops. Next after observing the certain columns it was observed that there were variations of the same value. Coming to our customer segmentation, the K-means algorithm works with only numeric data. Hence, a new pandas dataframe was created containing only our numeric columns. Finally, for our classification, the models worked well with numeric column hence our non-numeric categorical column were encoded.

The resulting dataset was then separated based on their clusters and within each cluster, the data was split into train and test with a proportion of 0.7 train data and 0.3 test data.

- **Models Used**

For customer segmentation, the K-mean clustering algorithm was implemented to create our customer groups. The following models were then compared to find the best fit for this study:

i. Support Vector Machine (SVM)

ii. Extreme Gradient Boosting Trees (XGBoost)

iii. Light Gradient Boosted Machine Algorithm (lightgbm) - proposed model.

- **Customer Segmentation**

Being inspired by [3], the customer segmentation was implemented using the K-means clustering algorithm.

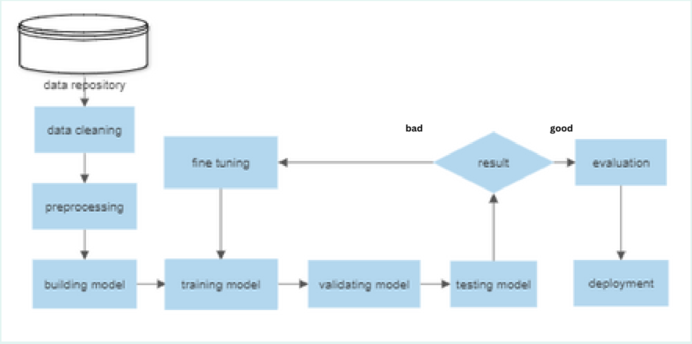
The clustering was done based on customer purchase behavior and the columns that were selected for this clustering where the Order Count column telling indicating the number of purchases the customer has made over the last month and OrderAmountHike FromLastYear indicating the increase in purchases compare to the previous year, expressed as a percentage. Out of this 3 clusters were formed.

- **Experimental Setup**

Here, the 3 models SVM, XGBoost and our proposed model , LGBM were trained on the various clusters of data and results were compared to find the best performing model, based on Accuracy and AUC/ROC metrics.

GENERAL OVERVIEW

Diagram of General Architecture



RESULTS

model	avg. accuracy	avg. ACC
svm	0.8836	0.6958
xgboost	0.9169	0.7882
LGBM	0.9132	0.9497

CONCLUSION

The results of this study also have some limitations. The study was carried out using an a single e-commerce customer data set with 5360 customers which may not reflex the entire behavior of customers world-wide. Ideally this research should have been verified with several data sets with a larger pool of customer data.

From experimental results, the proposed model LGBM was seen to observed to be the overall best during experimentation.

REFERENCES

[1] X. Yu, S.Guo, J. Guo, and X.Huang, “An extended support vector machine forecasting framework for customer churn in e-commerce”

[2] Shao, D. 2016. “Analysis and prediction of insurance company’s customer loss based on BP neural network. Lanzhou University”

[3] Xiahou, X.; Harada, Y. 2022. “B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM.” J. Theor. Appl. Electron. Commer. Res,17,458–475.