University of Ghana

LIGHT GRADIENT BOOSTING MACHINE FOR E-COMMERCE CUSTOMER CHURN PREDICTION

Presentation by

Vanessa Atta-Fynn

10665130

Computer Science | 2022

Overview

Abstract

Customer churn is what happens when the relationship of a customer with a company comes to the end. Losing customers is a serious problem that impacts all industries. Though e-commerce and retail services records one of the highest churn rates as shown by research carried out by Statista Research Department in 2020 on industry customer churn rates in the U.S, not much attention is paid to customer churn on e-commerce websites.

This study focuses on churn in the retail industry as it has one of the highest churn rates and traditional statistical analysis relies on high amount of assumption on customer data. In retail, this project will be narrowed further down to e-commerce as predicting churn in the offline setting will require more extensive data collection. In this project, machine learning algorithms will be used to predict customer churn in e-commerce.

Computer Science | 2022

Problem

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.
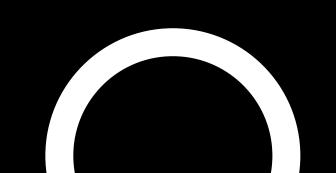
Contribution of Study

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.

Research Objectives

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.

Computer Science | 2022

Scope of Study

The scope of this study is in e-commerce as a representation of online retail with notable references dating up to present time of publication.

Limitations

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

To begin the experimental setup and study, a dataset with specified features was selected and extracted. Univariate, Bivariate and Mulitvariate Exporatory Data Analysis was performed on this data and problems identified were cleaned out. Feature Engineering was performed to prepare the data for the modelling. Customer Segmentation was then performed using the K-means clustering alorigthm and the models SVM, XGBoost and LGBM were trained and tested on the various segments. The results were compared to find the best performing model for e-cmmerce churn prediction using this dataset.

Computer Science | 2022

Data Cleaning

01 First missing values where identified and handled.

02 Irrelevant Columns with no contibution to study were identified and removed

03 Variation is column values were also identified and resolved.

04 Feature engineering was performed such as changing categorical columns to numeric.

Implementation

Phase 1

3 clusters were created using the K-means clustering algorithm. The clustering was done based on columns expressing customer purchase behavior.

Phase 2

The models, SVM, XGBoost and LGBM were built with the sci-kit learn library and trained on the 3 clusters created with the K-means algorithm

Phase 3

With the help of the sci-kit learn library, the SVM, XBoost and LGBM models were evaluated on using Accuracy, AUC and Precision Recall to find the best performing model

Computer Science | 2022

Result

At the end of the experimental study, it was observed that the Light Gradient Boosting Machine showed the overall best results with an accuracy of **0.9132** and AUC score of **0.9497.**

Model	Avg. Accuracy	Avg. AUC	TP Cluster (I+II+III)	FP Cluster (I+II+III)	TN Cluster (I+II+III)	FN Cluster (I+II+III)
SVM	0.8836	0.6958	1382	196	97	15
XGBoost	0.9169	0.7882	1365	124	169	32
LGBM	0.9132	0.9497	1386	43	241	11

Conclusion

With the help of the presented machine learning models and research performed, business owners and stakeholder in the e-commerce sector can now have the technical knowledge and tools to be able to easily adopt this technology into thier business setting to forecast likely churners and quickly take remedial action to improve customer satisfaction and retention.

Computer Science | 2022