Data Report on Customer Lifetime Value Predictions

Authors:

- Brenda Ngigi
- Beatrice Adhiambo
- Vincent Kiplang'at
- Brian Gikonyo
- Winnie Nzuve
- Cynthia Gitonga

Overview

Customer Lifetime Value (CLV) prediction is crucial for businesses offering subscription services. By accurately estimating the long-term value of customers, companies can tailor marketing strategies, improve customer retention efforts, and optimize resource allocation for maximum revenue. This proposal outlines a comprehensive approach to develop a CLV prediction model using machine learning techniques.

Business Understanding

In today's era of streaming services and a plethora of movie options, users often struggle to find movies that align with their preferences. A movie recommendation system addresses this challenge by leveraging user ratings to deliver tailored movie suggestions. Companies in the movie streaming industry, such as Netflix, Amazon Prime Video, and Hulu, can utilize recommendation systems to enhance customer engagement and retention. By offering personalized movie recommendations, these businesses can cater to diverse audience preferences, leading to growth and competitive advantage in the entertainment sector.

Objectives.

- Develop a customized movie recommendation system using cutting edge machine learning algorithms.
- Provide personalized movie recommendations for each user based on their preferences and viewing history.
- Enhance user experience and engagement by offering relevant and tailored movie suggestions.

Data Understanding.

Data Source.

The data set was obtained from Kaggle.

Data Description.

It has four documents; customer cases, customer info, customer product, and product info. It is later merged into one dataset.

The dataset consists of 330,512 entries and 22 columns capturing various aspects of customer interactions with subscription services. Key columns include unique identifiers for cases and customers, timestamps for interactions, demographic information such as age and gender, details about subscribed products and their pricing, as well as subscription specifics like billing cycles and signup/end dates. Additionally, the dataset contains information on customer lifetimes, including signup and cancellation dates, along with derived metrics such as lifetime months and customer lifetime value. These data elements provide valuable insights into customer behavior, subscription patterns, and long-term value, facilitating strategic decision-making for resource allocation, marketing optimization, and customer retention efforts in subscription-based businesses.

Data Preprocessing

1. Handle missing values, duplicates, and inconsistencies: in this step of the data analysis, the quality of the dataset is investigated and addressed. That requires that missing values, duplicates and other anomalies present within the dataset are handled accordingly.
   Missing values: In handling missing values, techniques like imputation and deletion are utilized. This step maintains the completeness and robustness of the analysis.
   Duplicates: Duplicate values bring about redundancy within the dataset. In removing these values the dataset is much more streamlined and thus much more meaningful insights can be obtained.
   Inconsistencies: Conflicts and erroneous entries are addressed in this step. In this stage of data cleansing and validation, the reliability of the dataset as well as its coherence is upheld.

2. Exploratory data analysis (EDA): in this stage, a deeper understanding to the idiosyncrasies of the dataset is established to understand data distribution and patterns.
   Patterns and Relationships: in the given dataset, the common feature is “movie id” from which all other data manipulation is performed. Other important features singled out are user ratings and genres.

Model Selection and Development

n the Model Selection section, after scaling the data using the Robust Scaler to mitigate the influence of outliers, we benchmarked our predictive modeling process with Linear Regression.

Following this, we explored more sophisticated algorithms, including XGBoost, Random Forest, and Neural Network, to capture complex relationships within the data. Through rigorous cross-validation, XGBoost emerged as the top-performing algorithm, exhibiting superior predictive capabilities and generalization performance.

Subsequently, we fine-tuned the hyperparameters of the XGBoost model using both Random Search and Grid Search techniques. Our experimentation revealed that Grid Search yielded the optimal hyperparameter configuration, resulting in improved predictive accuracy and stability.

The best-performing XGBoost model, with hyperparameters colsample_bytree=1.0, learning_rate=0.2, max_depth=5, n_estimators=300, and subsample=0.9, achieved remarkable performance metrics on the test dataset: RMSE of 0.0154, MAE of 0.00877, and an R^2 score of 0.9999999999491149.

Additionally, predictions on the test data using this tuned model yielded the following results: RMSE: 0.015363298481799917, MAE: 0.008768428199340126, and R^2: 0.9999999999491149.

Leveraging these findings, we deployed the optimized XGBoost model to predict customer lifetime values, empowering strategic decision-making and driving business growth.

Results

The analysis of key features impacting Customer Lifetime Value (CLV) revealed crucial insights for subscription-based businesses. From our investigation, two features emerged as significant determinants of CLV:

1. Lifetime Months: The duration of a customer's subscription tenure, represented by Lifetime Months, was found to strongly influence CLV. Notably, the type of subscription (annual or monthly) exerted a notable impact on this feature. Customers with annual subscriptions demonstrated a longer subscription

lifetime compared to their monthly counterparts, indicating the importance of subscription duration in driving CLV.
2. Price: The price of the subscribed product emerged as another critical factor affecting CLV. Higher-priced subscriptions exhibited a positive correlation with higher CLV, suggesting that customers investing more in their subscriptions tend to hold greater long-term value to the business.

Our visualization of the distribution of Lifetime Months against Customer Lifetime Value underscored the significance of customer retention strategies in maximizing CLV and long-term profitability. The observed trend indicates that customers with extended subscription lifetimes tend to possess higher CLV, highlighting the imperative for businesses to prioritize retention efforts to foster loyalty and maximize CLV.

Conclusion

Our rigorous modeling efforts aimed at predicting Customer Lifetime Value (CLV) have provided invaluable insights for subscription-based businesses. Through meticulous data preprocessing, feature engineering, and model selection, we identified XGBoost as the optimal algorithm, delivering exceptional predictive accuracy and performance metrics.

Key findings underscore the significant impact of two primary drivers on CLV: the duration of customer subscription tenure and the pricing of subscriptions. Notably, the type of subscription (annual or monthly) was found to influence subscription duration and consequently CLV.

Our visualization of the relationship between Lifetime Months and Customer Lifetime Value highlights the critical importance of customer retention strategies in maximizing CLV and long-term profitability. By leveraging our optimized XGBoost model, businesses can drive strategic decision-making, enhance marketing efforts, and optimize resource allocation to foster customer loyalty and drive sustainable growth in the competitive subscription services landscape.

Limitations

- Limited Models Used: The CLV model's performance is constrained by the selection of machine learning algorithms considered during the modeling

process. Additional algorithms or ensemble methods might offer different insights and improve predictive accuracy.

- Model Generalization: While the CLV model may perform well on the training and test datasets, its performance on unseen data or in different market contexts may vary. Generalizing the model to diverse customer segments and market conditions requires careful validation and testing.
- Interpretability: Complex machine learning models like XGBoost may lack interpretability, making it challenging to understand the underlying factors driving CLV predictions. Clear interpretation of model outputs is essential for effective decision-making.