kaggle

KaggleX- Showcase

Customer Lifetime Value Prediction in Retail Industry

Presentation Title

Purity Supaki

Background

- I am a Data scientist at a fintech company. I have 4 years of experience in data analysis, machine learning, and statistical modeling. In my current role, I am working on projects that involve building predictive models to solve business problems and improve decision-making. I am passionate about using data science to drive innovation and create value for organizations.
- Currently pursuing a M.Sc. Artificial Intelligence
- My goal is to apply data science techniques to solve real-world business problems and drive meaningful impact in the fintech and retail space.
- Project presentation is on customer lifetime value prediction which is interesting to me as I believe it has the potential to revolutionize the way retailers approach customer relationship management and drive long-term growth. It also relates to my goal of using data science to solve real-world problems in the retail industry.

PROJECT SUMMARY

Project Summary

- Customer lifetime value (CLV) is a crucial metric for any retail company. It represents the total amount of revenue a customer is expected to generate during their lifetime with the company, taking into account factors such as purchase frequency, average order value, and customer retention rates.
  - A higher CLV indicates a more loyal and valuable customer, which translates into increased profitability for the company.
  - A high CLV indicates a loyal and valuable customer who is likely to make repeat purchases and recommend the company to others.
- The goal of my project is to build a predictive model that estimates CLV for customers in a retail company. This model will use historical transaction data to predict future revenue potential and help the company make data-driven decisions around marketing, promotions, and customer retention. By accurately predicting CLV, the company can tailor its marketing strategies, improve customer retention, and increase revenue growth.

Project Details

Customer lifetime value helps answer some questions that are critical in determining future strategic decisions:

- How much do you need to invest to acquire, engage, and retain your customers?
- Does the cost of maintaining your relationship with your customers exceed their CLV?
- Which of your products and services have the highest profitability?
- What is the financial impact of your business strategies and marketing efforts?

Benefits of CLV:

- Improved customer loyalty: Monitoring clv of customers helps in identifying the best customers and determining marketing strategies that work for them to retain them.
- Increased repeat sale: Increasing frequency of transactions of a user leads to higher CLV
- Increased profitability: Higher clv leads to higher profits

Data Science Topic(s) Applied

- To build an accurate CLV model, I applied several data science topics and techniques, including data cleaning, exploratory data analysis, feature engineering, model selection, and evaluation.
- In this project, I applied several data science topics and techniques, including data cleaning, exploratory data analysis, feature engineering, model selection, and evaluation.
- Feature engineering was a critical step in this project. I created several new features, such as recency, recency clusters, frequency, frequency cluster and monetary value (revenue), that capture important aspects of customer behavior .
- These topics are relevant for building a CLV model because they help to identify and capture the key factors that drive customer behavior and purchasing patterns. For example, data cleaning ensures that the data is accurate and complete, exploratory data analysis helps to identify patterns and trends in the data, and feature engineering creates new variables that capture important information about customer behavior.

Data Collection and Preprocessing

- For this project, I collected transaction data from a retail company's database. The transaction data included the date of the purchase, the product SKU, the quantity purchased, and the total amount spent and purchase history.
- I cleaned and preprocessed the data by dealing with missing values, outlier detection, and data transformation.
- To illustrate my approach, I created some visualizations of the cleaned data, such as histograms, scatter plots, and box plots. These visualizations helped me to identify patterns in the data and confirm my assumptions about customer behavior.

Feature Engineering

- Feature engineering is the process of creating new variables that capture relevant information from the original data. In this project, I engineered several features for my CLV model, such as recency, frequency and monetary value, to capture important aspects of customer behavior and purchasing patterns, including:
  - Recency: the number of days since the customer's last purchase; measures how recently a customer made a purchase
  - Frequency: the number of purchases the customer has made in the past year; measures how often they make purchases
  - Monetary value (Revenue): the total amount the customer has spent with the company in the past year
- These features were selected based on their relevance to predicting CLV and were derived from the transaction and customer data. Customer demographics, such as age, gender, and income, can also be useful predictors of CLV.

Model Selection and Evaluation

- To predict CLV, I considered several types of models, including linear regression, decision trees, and XGBoost. I evaluated the models based on criteria such as mean squared error, R-squared, and accuracy.

- The XGBoost model performed the best, with an accuracy of 94% on the test set. The model was able to accurately predict CLV for high, mid, and low-value customers, as indicated by the precision, recall, and f1-score metrics in the classification report.

Model Deployment & Future Work

- To deploy the CLV model in a real-world setting, I would integrate it with a CRM system or a marketing automation platform. This would allow the company to segment customers based on their CLV and tailor marketing campaigns accordingly.
- However, there are limitations to the current model that need to be addressed. For example, incorporating more data sources, testing different model architectures, or using deep learning techniques could improve the accuracy and for model deployment, we would recommend integrating the CLV model with existing customer databases and marketing systems. This would enable marketers to target high-value customers with tailored promotions and offers, as well as identify at-risk customers who may require special attention. However, it is important to keep in mind that deploying machine learning models in a real-world setting requires careful consideration of data privacy, security, and ethical issues.
- As for future work, there are several avenues for improving the CLV model. One possibility is to incorporate additional data sources, such as social media activity, customer reviews, or website engagement metrics, which may provide more insights into customer behavior and preferences. Another option is to experiment with different modeling techniques, such as ensemble models, deep learning algorithms, or reinforcement learning frameworks, to see if they can further improve the accuracy and robustness of the model. Lastly, we could explore the impact of different marketing strategies on CLV, such as A/B testing, personalized messaging, or loyalty programs, to better understand how to optimize customer lifetime value.

What I Learned

- In this project, I learned several key lessons about data science and machine learning. First, I discovered the importance of data cleaning and preprocessing, which can significantly impact the performance and accuracy of a model. Second, I realized that feature engineering is a crucial step in building a predictive model, as it allows us to extract meaningful insights from the data and create new variables that capture important aspects of customer behavior. Third, I gained insights into the trade-offs between different modeling techniques, such as linear regression, decision trees, and XGBoost, and learned how to evaluate their performance using metrics such as mean squared error, R-squared, and accuracy.

- Overall, this project has helped me develop my skills in data science and machine learning, and has given me a better understanding of how these techniques can be applied to real-world problems in the retail industry. I look forward to applying these skills in future projects and continuing to explore the exciting and rapidly evolving field of data science kaggle