Data Science
Final Project

NAVIGATE E-COMMERCE THROUGH PREDICTIVE CHURN ANALYSIS AND CUSTOMER SEGMENTATION

by: Shofi Dh

June 2024

OUTLINE

01 **Introduction**
- Background
- Objective
- Scope

02 **Business & Data Understanding**
- Data Understanding

03 **Data Preparation**
- Data Splitting
- Handling missing values and duplicates
- Exploratory Data Analysis
- Feature Engineering (Encoding & adding RFM feature)

04 **Churn Prediction**
- Machine Learning Model
- Model Evaluation
- Hyperparameter Tuning

05 **Customer Segmentation**
- RFM Analysis
- Segmentation Classification
- Segmentation Evaluation
- Segmentation Profiling

06 **Conclusion & Business Insight Recommendation**

07 **Benefit for the Company**

08 **Reference**

INTRODUCTION

89Stocker

In the dynamic e-commerce landscape, where **retaining customers is pivotal yet challenging, existing customers offer significant value while attracting new ones demands substantial investments** (Wu et al., 2017). This study aims to develop a predictive model tailored to the e-commerce sector, correlating key attributes with churn, to improve customer retention and enhance marketing strategies. Collaborating with customer segmentation, the study seeks to **understand customer behavior, reduce churn, and segment customers for targeted marketing efforts.**

Source: Wu, X. J., & Meng, S. S. (2017). Research on e-commerce customer churn prediction based on customer segmentation and Ada-Boost. Industrial Engineering, 20(02), 99-107.

BUSINESS & DATA UNDERSTANDING

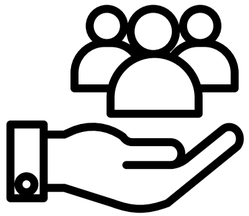
Data Understanding

This project uses the dataset "E-commerce Customer Behavior and Purchase Dataset" from Kaggle, which was synthesized using the Faker Python library (license: Open Data Commons).

25.000
Rows

Transaction Data:
- Purchase Date
- Product Category
- Product Price
- Quantity
- Total Purchase Amount
- Payment Method
- Returns

12
Features



Customer Demographics:
- Customer Name
- Customer Age
- Gender

1
Target

Target variable
- Churn

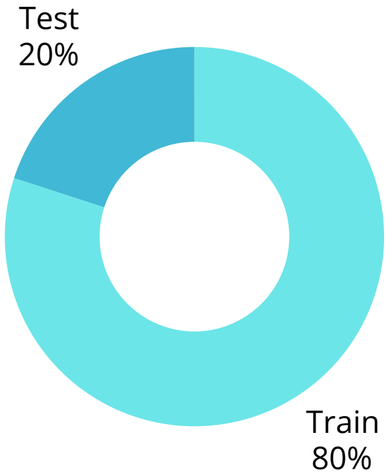
Business Goals:

1. Understand customer behavior,
2. Reduce churn, and
3. Segment customers for targeted marketing efforts.

DATA PREPARATION

A. Data Spliting

Data split is done at the biginning, the goal is to avoid data leakage.



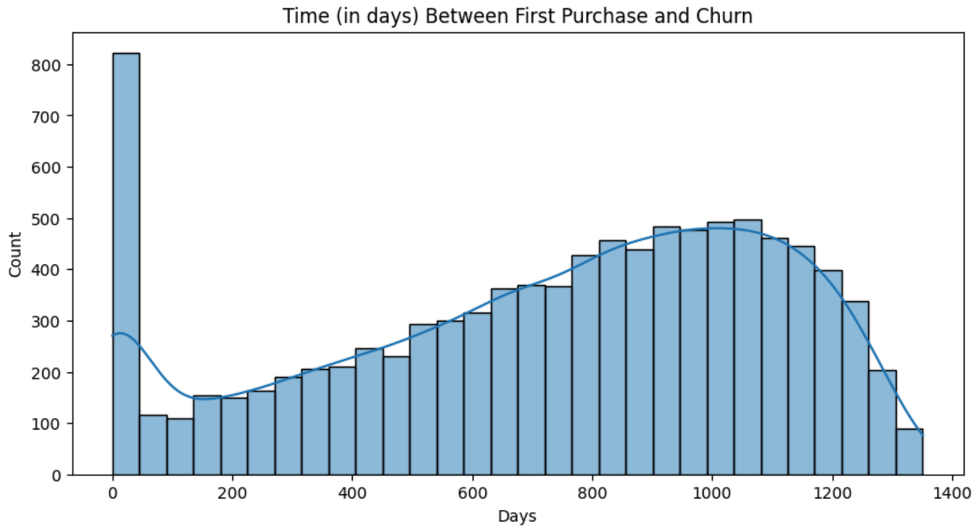
B. Data Cleaning

19% (train) &
18.9% (test)
Missing Value
Returns coloumn

0
Duplicated Data

C. Exploratory Data Analysis

By **doing EDA only on the training set**, the model is less likely to be influenced by patterns or outliers from testing data set.



There appears to be significant customer churn within the first 30 days of usage and also after 1000 days of usage.

C. Exploratory Data Analysis

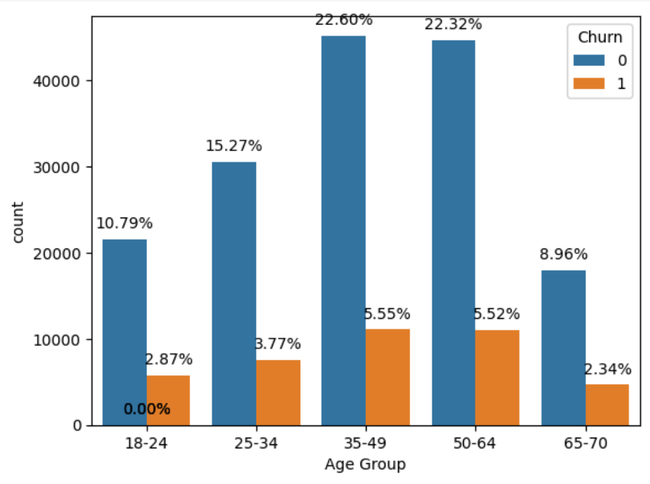
Chi-square & p-value test features and churn

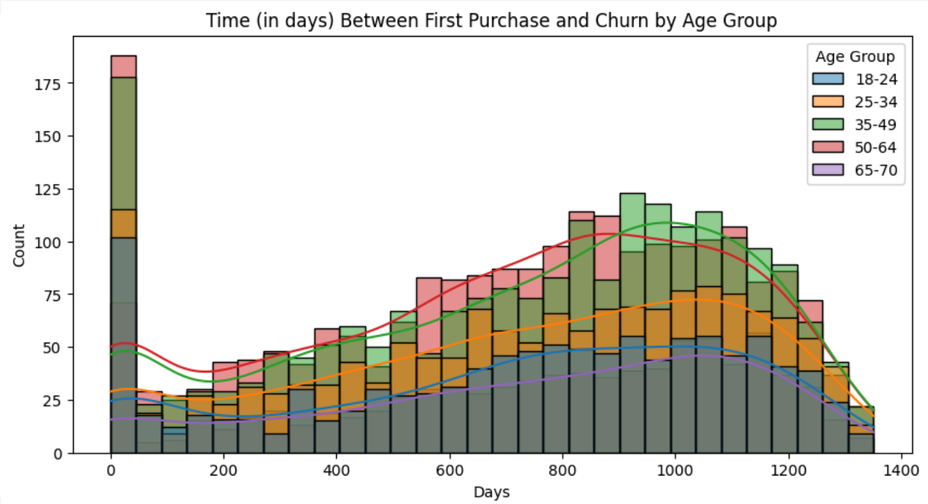
Variable	Chi-square valuee	P-value	Conclusion
Product Category	1.906	0.592	Independent, Accept H0
Payment Method	7.263	0.026	not independent, Reject H0
Gender	1.475	0.225	Independent, Accept H0
Age Group	29.546	6.055e-06	not independent, Reject H0
Returns	3.292	0.070	Independent, Reject H0

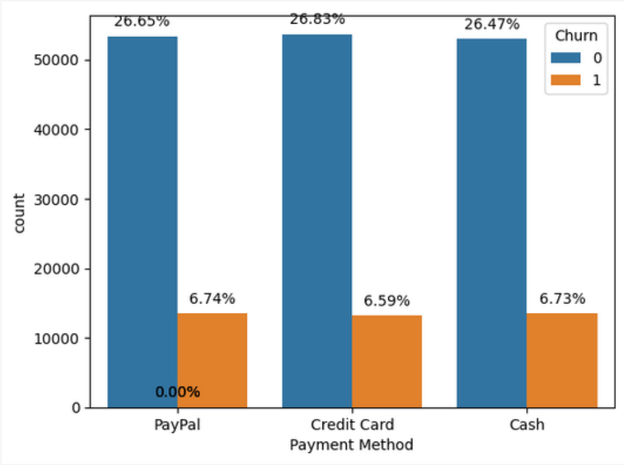
- Focus more on Payment Method and Age Group as they show significant associations with Churn.
- Product Category, Gender, and Returns do not seem to have significant associations with Churn based on these tests.

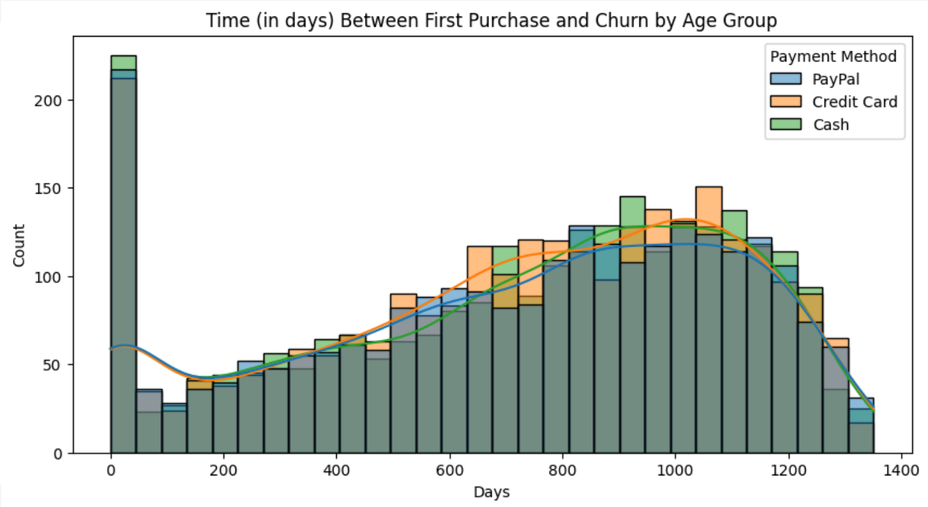
Variable	Churn Rate (%)
Payment Method	
Cash	**20,26 %**
Credit Card	19,71 %
PayPal	20,17 %
Age Group	
18-24	**21,02%**
25-34	19,82%
35-49	19,17%
50-64	19,82%
65-70	20,70%

C. Exploratory Data Analysis









89Stocker

D. Feature Engineering

Activity	Train & Test Respective data sets
1. Handling Outlier	**Handling outlier using IQR in Total Purchase Amount coloumn**
2. Handling Missing Value	**Using mode imputation to handling missing value in Returns coloumn**
3. Encode Dataset	**Encode data categorical like: Product Category, Payment Methode, Gender, Purchase Month and Age Group**
4. Feature Scalling	**Separating columns into those requiring standardization for numerical data and those that do not need standardization for categorical data.**

CHURN PREDICTION

Machine Learning Model	F1 Score	
	Baseline	SMOTE
Logistic Regression	0.000	25,9%
KNeighbors Classifier	8,5 %	26,6 %
GaussianNB	0.000	25,4 %
Decision Tree Classifier	21,5 %	24,0 %
Random Forest Classifier	0,6 %	16,0 %
XGB Classifier	0,01%	14,9 %

Hyperparameter Tuning: KNeighbors Classifier

Best Parameters:
{'algorithm': 'ball_tree',
'leaf_size': 47,
'metric': 'manhattan',
'n_jobs': None,
'n_neighbors': 1,
'p': 2,
'weights': 'distance'}

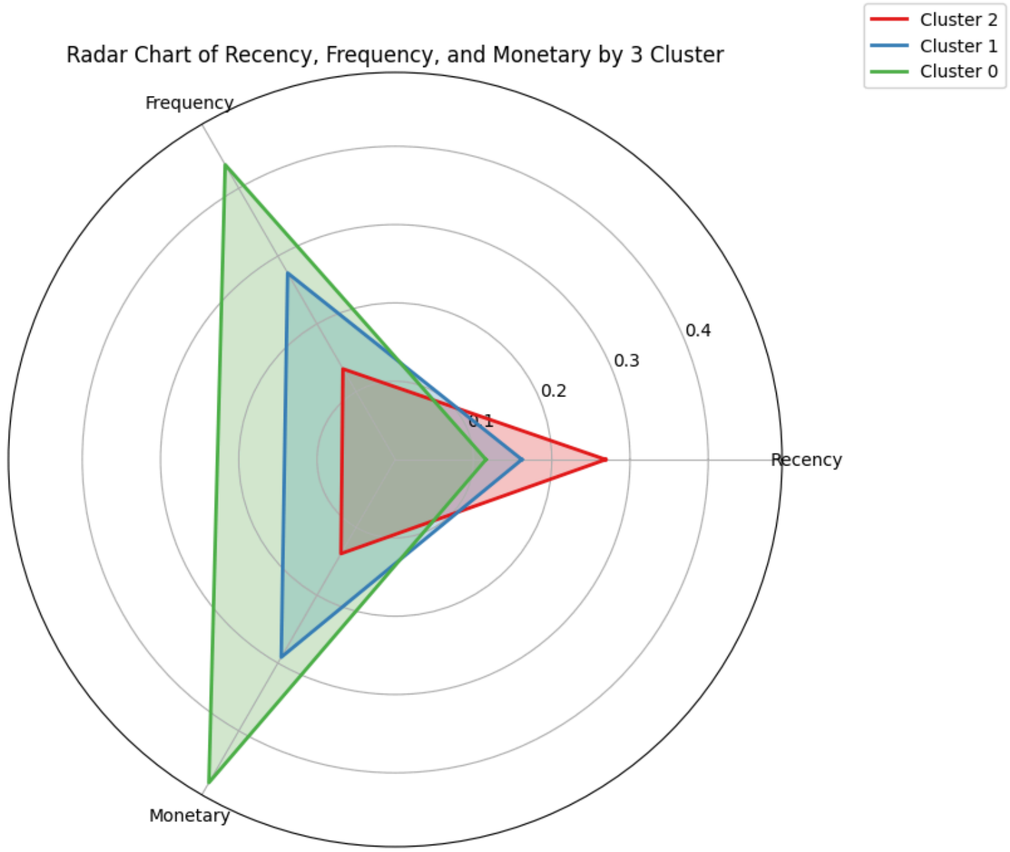
Best F1 Score: 82,02%

CUSTOMER SEGMENTATION

A. RFM Analysis

Variable	Tabel
Recency	calculated as the number of days between the customer's last purchase date and the latest purchase date recorded in the dataset (Max Purchase date subtracted by last purchase date).
Frequency	The total number of purchases made by the customer (size of Purchase Date grouped by Customer ID).
Monetary	The total monetary value of all purchases made by the customer (sum of Total Purchase Amount grouped by Customer ID).

B. Segmentation Clasification



- **Cluster 0: High-Value Customers**: These customers recently purchased with a mean recency of 156.58 days, averaging 7.96 purchases, and a mean monetary value of $24,196.46.
  [**Recency**, **Frequency**, **Monetary**]

Focus on retention through engagement, offer upselling/cross-selling opportunities, and gather feedback regularly

- **Cluster 1: At-Risk Customers**: With a mean recency of 218.65 days, these customers average 5.40 purchases and have a mean monetary value of $14,853.70.
  [**Recency, Frequency, Monetary**]

Implement re-engagement strategies, emphasize value propositions, and conduct surveys to address concerns.

- **Cluster 2: Dormant Customers**: This cluster has the highest recency (mean: 363.45 days), lowest purchase frequency (mean: 3.14), and lowest monetary value (mean: $7,135.09).
  [**Recency**, **Frequency**, **Monetary**]

Focus on reactivation with compelling incentives, run win-back campaigns, and segment further for tailored marketing efforts.

HIghest
Middle
Lowest

B. Segmentation Evaluation & Profiling

Cluster	Number of Costumer	Churn Percentage (%)	Churn Rate (%)
Cluster 0: High-Value Customers	9493	3,85 %	20,18 %
Cluster 1: At-Risk Customers	21383	8,55 %	19,86 %
Cluster 2: Dormant Customers	18785	7,58 %	20,04 %

Retention Focus on Cluster 1

It is the largest but with the highest churn percentage, may require immediate attention in terms of retention strategies. Understanding the reasons behind churn in this cluster and implementing targeted retention campaigns could help mitigate further losses.

CONCLUSION & BUSINESS INSIGHT RECOMMENDATION

Exploratory Data Analysis

1. The analysis indicates that Payment Method and Age Group are significant factors associated with customer churn, it's essential to tailor strategies based on these characteristics.
2. Payment Method Recommendation: Offer discounts or cash back for using preferred payment methods and Reward loyalty points for using certain payment options.
3. Age Group Recommendation, use engagement tactics:

- Younger Audience (18-24): Utilize social media platforms like Instagram and TikTok, Offer student discounts and relevant promotions.
- Middle Age Group (25-49): Highlight convenience and family-oriented products & Promote work-life balance and self-care items.
- Older Audience (49+): Emphasize quality and reliability & ensure a user-friendly website and excellent customer support.

CONCLUSION & BUSINESS INSIGHT RECOMMENDATION

Churn Prediction

1. Model Performance:

- The baseline F1 scores for most models were significantly low, indicating poor predictive performance.
- After applying SMOTE (Synthetic Minority Over-sampling Technique), F1 scores improved across most models, demonstrating the effectiveness of addressing class imbalance.

2. Best Performing Model:

- **The KNeighbors Classifier, after hyperparameter tuning, achieved the highest F1 score of 82.02%.**
- This indicates that the KNeighbors Classifier, with optimized parameters, is the most effective model for predicting churn in this dataset.

CONCLUSION & BUSINESS INSIGHT RECOMMENDATION

RFM Analysis

1. Cluster 0 High-Value Customers: Focus on retention through engagement, offer upselling/cross-selling opportunities, and gather feedback regularly
2. Cluster 1 At-Risk Customers: Implement re-engagement strategies, emphasize value propositions, and conduct surveys to address concerns (**Retention Focus on Cluster 1,** It is the largest but with the highest churn percentage).
3. Cluster 2 Dormant Customers Focus on reactivation with compelling incentives, run win-back campaigns, and segment further for tailored marketing efforts.

BENEFIT FOR THE COMPANY

I assume the company has 1,000 customers, an average customer acquisition cost (CAC) of IDR 1,000,000, and an average annual revenue per customer of IDR 5,000,000.

1. Improved Customer Retention

- **Metric**: Increase in customer retention rate by 10% over the next 12 months.
- **Assumption**: By predicting likely churners with 80% accuracy, the company can engage 50% of these customers with targeted strategies (e.g., personalized offers, discounts, enhanced support).
- **Impact**: Retaining an additional 50 customers per year.
- **Financial Impact: 50 customers * IDR 5,000,000 = IDR 250,000,000 in retained revenue annually.**

2. Cost Efficiency

- **Metric**: Reduction in customer acquisition cost (CAC) by 15% within 6 months.
- **Assumption**: Retaining customers costs 50% less than acquiring new ones, resulting in significant cost savings.
- **Impact**: Saving approximately IDR 750,000 on acquisition costs per retained customer.
- **Financial Impact: 50 customers * IDR 750,000 = IDR 37,500,000 saved annually on acquisition costs.**

3. Revenue Growth

- **Metric**: 5% increase in annual revenue due to reduced churn.
- **Assumption**: Satisfied customers are 2x more likely to continue using services and make additional purchases.
- **Impact**: Generating an additional 5% revenue from the existing customer base.
- **Financial Impact: 1,000 customers * IDR 5,000,000 * 5% = IDR 250,000,000 in additional annual revenue.**

BENEFIT FOR THE COMPANY

Summary of Financial Impact

- Total Retained Revenue. **IDR 250,000,000**

- Cost Savings on Acquisition. **IDR 37,500,000**

- Additional Revenue from Growth **IDR 250,000,000** **+**

Total Annual Financial Impact **IDR 537,500,000**

- By focusing on these measurable benefits and making realistic assumptions, the company can clearly see the value of implementing strategies to reduce churn and improve customer retention, resulting in a significant financial impact of IDR 537,500,000 annually.

Data Science
Final Project

REFERENCE

- Wu, X. J., & Meng, S. S. (2017). Research on e-commerce customer churn prediction based on customer segmentation and Ada-Boost. Industrial Engineering, 20(02), 99- 107.
- https://towardsdatascience.com/two-rookie-mistakes-i-made-in-machine-learning-improper-data-splitting-and-data-leakage-3e33a99560ea
- https://medium.com/@madhuri15/the-roles-of-data-sets-in-machine-learning-projects-a-guide-to-data-splitting-454beb468a49

THANK YOU

FOR YOUR NICE ATTENTION

+62851-7172-7676

shofidh.contact@gmail.com

https://www.linkedin.com/in/shofidh/

Bandung Regency, 40228

click here to check the code and the whole process

June 2024