**Customer segmentation with PyCaret**

Table of content

1. Introduction
2. Data overview
3. Exploratory Data Analysis
4. Data cleaning
5. Descriptive Data Analysis
6. Diagnostic Data Analysis
7. Data preprocessing
8. Customer Segmentation
9. **Introduction**

In this project, I will perform unsupervised clustering of customer data from a bank with PyCaret. Customer segmentation is the process of grouping customers who share certain traits or characteristics into segments to optimize marketing strategy, maximize customer value to the business, and improve customer experience and satisfaction.

1. **Data overview**

The original dataset contains 113066 data points and 10 attributes. The attributes are as followed:

* customer\_id: Customer id number
* prod\_ca: Payment account
* prod\_td: Time deposit
* prod\_credit\_card: Credit card
* prod\_app: Banking mobile app
* prod\_secured\_loan: Secured loan
* prod\_upp: Unsecured loan
* amount: Account balance
* segment: Customer segment
* province\_city: Residential city

Graphical user interface, application

Description automatically generated

Figure 2.1 First 5 rows of the dataset.

1. **Exploratory Data Analysis**

At first glance, the dataset contains information about 113066 customers from 42 regions, which are segmented into 3 groups.

Timeline

Description automatically generated with medium confidence

Figure 3.1 Total customers in each region.

Graphical user interface

Description automatically generated

Figure 3.2 Total customers in each segment.

Using pandas-profiling for EDA, I receive following alerts:

* The amount column is highly skewed, so I need to explore this column in order to identify the risk of outliers.
* There are 10144 zeros in the amount column so I will remove those rows as the amount is an important feature of the clustering model.

Application

Description automatically generated with low confidence

Figure 3.3 Alerts from pandas-profiling.

Exploring the amount column, I detect outliers in the amount column. So, I will use the IQR method to reduce outliers in the Data preprocessing section.

Text

Description automatically generated with medium confidence

Figure 3.4 Exploring amount column with describe().

Chart

Description automatically generated

Figure 3.5 Exploring amount column with box plot.

1. **Data cleaning**

From pandas-profiling alert, I discover:

* There are 996 missing values in prod\_credit\_card. Since prod\_credit\_card is a categorical variable, I replace missing values with 0, which indicates not using credit card service.
* 9% of the users (10144 rows) have 0 account balances.

After replacing missing values in prod\_credit\_card and removing rows with 0 account balance, the total number of data points after cleaning is **102922**.

1. **Descriptive Data Analysis**
2. Province city analysis

* Ho Chi Minh and Ha Noi are the two cities with the most customers. Together, these two accounts for more than half of the total customers. This is understandable as these are the largest cities in Vietnam.
* Noticeably, there are nearly 5 thousand customers without information about province city.

Chart, pie chart

Description automatically generated

Figure 5.1 Total customers by city.

1. Regional analysis

* The North and the South accounts for 85% of the total customer. However, the Middle region only contributed 10,000 customers to the total. This figure is very humble considering the potential in this market. The company should invest more human resources and budget into expanding business activities in this region.

Chart, pie chart

Description automatically generated

Figure 5.2 Total customers by region.

* Having the most customers, the North and the South also lead in total product usage and total amount. However, the total bank amount of the North is only half of the South’s. I will look further into the average bank amount to find the reason for this matter.

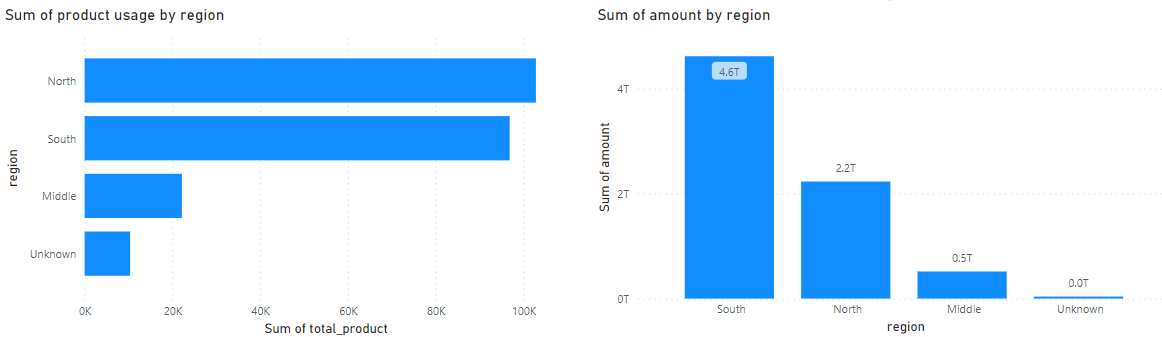


Figure 5.3 Total product usage and total account balance by region.

* Though having the greatest number of customers, the North region fails to convert it into product usage and bank deposit. To raise product usage rate and average deposit amount, branches in the North region should change their way of approaching, providing discounts or preparing more attractive deposit plans for the customers.

Chart, line chart

Description automatically generated

Figure 5.4 Average amount and average product usage by region.

1. Segment analysis

* The regular segment is the most common one when 80% of the customers are in this segment. This is understandable as there should be many more regular customers than loyal customers.

Chart, pie chart

Description automatically generated

Figure 5.5 Total customers by segment.

* Having more customers, the Regular segment has the most total product usage with about 175,000. However, the Gold segment has almost 6 trillion bank deposits, compared to under 1 trillion for Regular and Silver each.

Graphical user interface, chart, application

Description automatically generated

Figure 5.6 Total product usage and total account balance by segment

* The average amount and average product usage explain how the segment system works. Gold customers are those with high deposit amounts and product usage and Silver customers are those with middle product usage.

Chart, line chart

Description automatically generated

Figure 5.7 Average account balance and product usage by segment

1. **Diagnostic Data Analysis**
2. What is the most used product?

* As online payment is used widely in Vietnam, banking apps and payment accounts are the two most used.
* Secured loans and unsecured loans are the least used product, this may come from the fear of being in debt of Vietnamese.

Chart, bar chart

Description automatically generated

Figure 6.1 Total usage per product.

1. Is account balance and product usage correlated?

* Customers with higher account balance have a slight tendency to use more service from the bank.

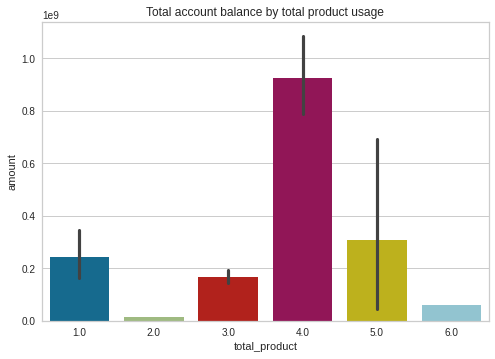


Figure 6.2

Chart, bar chart

Description automatically generated

1. **Data preprocessing**

After validating datasets, I will preprocess data with the following steps:

* Categorize province\_city column into regions: North, Middle, and South
* Using the IQR method to remove outliers in the amount column: **20690** outliers are detected and removed.
* Label encoding the categorical features.

After preprocessing, data includes 82232 data points.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 7.1 First 5 rows of the dataset

1. **Customer Segmentation**

PyCaret is an open-source, low-code machine learning library in Python that automates machine learning workflows. With PyCaret, I will segment customers into groups.

The attributes I will use to train the model are prod\_ca, prod\_td, prod\_credit\_card, prod\_app, prod\_secured\_loan, prod\_upl, amount, total\_product, segment\_encode, and region\_encode.

Graphical user interface, application

Description automatically generated

Figure 8.1 Top 5 rows of training dataset

The model chooses the number of clusters as 4.

Chart, line chart

Description automatically generated

Figure 8.2 Elbow score

Cluster pattern:

* Cluster 0: Low account balance and high total product usage.
* Cluster 1: Low account balance and low total product usage
* Cluster 2: Low account balance and common total product usage
* Cluster 3: High account balance and high total product usage

Calendar

Description automatically generated with medium confidence

Figure 8.3 Product usage and account balance per cluster

Chart, bar chart

Description automatically generated

Figure 8.4 Total customers per cluster