

CHAPTER 4 – RESULTS AND DISCUSSION

This chapter explores into the data analysis of PT XYZ, a large Internet Service Provider in Indonesia, with an emphasis on customer churn and marketing intelligence. The research attempts to find trends and causes leading to customer churn using a Multilayer Perceptron (MLP) neural network and Exploratory Data Analysis (EDA), as well as significant marketing insights. The information, which is filled with customer-related variables, serves as a solid basis for both predictive modeling and strategic marketing research.

4.1 Data Overview

The PT XYZ dataset emerges as a valuable reservoir of information for studying consumer behavior in PT XYZ. It has 84,605 entries and a varied spectrum of 36 characteristics, each providing unique insights.

Table 4.1 PT XYZ Details Dataset Information

Column Name	Details Information
<i>prepaidType</i>	type of service
<i>Customer_Name</i>	Customer name
<i>tik_id</i>	Ticket ID for retention
<i>tik_Custid</i>	Customer identification (CID) number
<i>tik_CA_Number</i>	Customer Account (CA) number
<i>SIA_Number</i>	Service Instant Account number
<i>BA_number</i>	Billing Account number
<i>cat1</i>	Ticket closing category that shows about customer churn or not
<i>sub1</i>	The derivative of cat1
<i>seg</i>	Derivatives of sub1
<i>closeDate</i>	Retention ticket closing date
<i>closeTime</i>	Retention ticket closing time
<i>dayPart</i>	The length of day the retention ticket opened
<i>monthPart</i>	Length of month the retention ticket be opened
<i>typeRetentionDay</i>	retention day type
<i>typeRetentionMonth</i>	retention month type
<i>PACKAGEPRICE_IDR</i>	Price of service package
<i>PROFIT_CENTER_NAME</i>	Branch name
<i>Mapping_Area</i>	Customer area mapping
<i>Technology</i>	Fiber optic technology
<i>SA_Province</i>	Province
<i>trendcall</i>	Type of call
<i>DATE_ACTIVATION</i>	CID activation date
<i>DATE_EXPIRED</i>	CID expiration date
<i>PACKAGENAME</i>	Package name
<i>territory</i>	Territory area
<i>profit_center</i>	Branch code
<i>customergender</i>	Customer gender
<i>CITY</i>	City
<i>DISTRICT</i>	District
<i>SUB_DISTRICT</i>	Sub district
<i>Total Ticket</i>	Total tickets from customers
<i>SUBSCRIPTION_AGE</i>	Customer tenure
<i>CHURN</i>	Churn category
<i>DOB</i>	date of birth
<i>CSTMAGE</i>	customer age

The dataset consists of both category and numerical variables. *customergender*, *prepaidType*, and *CITY* are examples of categorical variables which offer qualitative insights. While *SUBSCRIPTION_AGE* and *CHURN* are examples of numerical variables that provide quantifiable measures. Their interaction is critical in understanding the complexities of client turnover, which is the key in this research. Each element serves as an individual piece, creating a complete picture of customer interactions and trends, which is critical for anticipating churn and developing successful retention efforts.

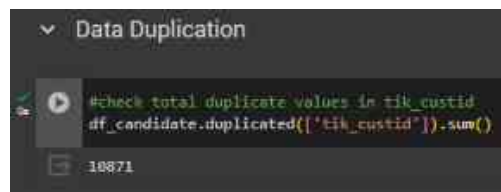
4.2 Data Preparation and Preprocessing

Upon analyzing and understanding the dataset, the initial step is to standardize the variables/features into lowercase due to the case sensitivity of Python programming. After completing feature standardization, the next step involves selecting which variables or features to use. Conducting feature selection reduces computational time, improves prediction performance, identifies relevant features, and enhances data quality (Awan et al., 2019). The features were simplified to a set of 23, with details provided in Table 4.2.

Table 4.2 Standardize Feature and Feature Selection Result

Category	Feature
Demographic Information	customergender, cstmr_age
Subscription Details	packageprice_idr, subscription_age, packagename, date_activation, date_expired
Billing and Payment Information	profit_center_name
Communication and Service Interaction	cat1, sub1, seg, closedate, total ticket
Geographic Information	mapping_area, sa_province, territory, city, district, sub_district
Technology and Network	technology
Customer Account Information	tik_custid, prepaidtype
Target Variable	churn

Next, the process will involve checking and ensuring data duplication is handled appropriately. This step is crucial, as careless elimination of duplicate data can lead to the loss of important information. The variable selected for checking data duplication is *tik_custid*, chosen because it is a unique identifier for each customer, with one customer having only one ID. As per Figure 4.1, there are 10,871 duplicate data entries based on customer ID. Care is required in this process due to the significant volume of data, which could impact the model, particularly the 'total ticket' feature. To address this, a new feature named *total_tickets* will be created. This step will consolidate ticket information based on customer ID into one row, ensuring no data loss. To avoid confusion, the original 'total ticket' feature will be dropped, as it is replaced by the new *total_tickets* feature. Consequently, the dataset will comprise 73,734 entries with the new feature *total_tickets*, as illustrated in Figure 4.2.



```
#check total duplicate values in tik_custid
df_candidate.duplicated(['tik_custid']).sum()
```

10871

Figure 4.1 Data Duplication in Dataset



```
df_candidate.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 73734 entries, 301 to 80501
Data columns (total 23 columns):
 #   Column                Non-Null Count  Dtype  
---  --
 0   customergender        73734 non-null  object  
 1   cstmr_age             73734 non-null  int64   
 2   tik_custid            73734 non-null  object  
 3   prepaidtype           73734 non-null  object  
 4   packageprice_idr      73734 non-null  int64   
 5   subscription_age      73734 non-null  float64  
 6   packagename           73734 non-null  object  
 7   date_activation       73734 non-null  object  
 8   date_expired          73734 non-null  object  
 9   cat1                  73734 non-null  object  
10   sub1                  73731 non-null  object  
11   seg                   73712 non-null  object  
12   closedate             73734 non-null  object  
13   profit_center_name    73734 non-null  object  
14   mapping_area          73734 non-null  object  
15   technology            73734 non-null  object  
16   sa_province           73734 non-null  object  
17   territory             73734 non-null  object  
18   city                  73734 non-null  object  
19   district              73734 non-null  object  
20   sub_district          73734 non-null  object  
21   churn                 73734 non-null  int64   
22   total_tickets         73734 non-null  int64   
dtypes: float64(1), int64(4), object(18)
memory usage: 13.5+ MB
```

Figure 4.2 Data Duplication Result and New Feature

Following the processing of duplicate data, the next step is to address the missing values in the dataset. As indicated in Figure 4.3, there are missing values in the features *sub1* and *seg*. Although the number of entries with missing values is not substantial, the decision was made not to drop these entries, recognizing that both *sub1* and *seg* are related to communication interaction, which is crucial to the analysis.



```
# finding nan value
df_candidate.isna().sum()

customergender      0
cstar_age           0
tik_custid          0
prepaidtype         0
packageprice_idr    0
subscription_age    0
packagename         0
date_activation     0
date_expired        0
cat1                0
sub1                 3
seg                 22
closedate           0
profit_center_name  0
mapping_area        0
technology          0
sa_province         0
territory           0
city               0
district            0
sub_district        0
churn               0
total_tickets       0
dtype: Int64
```

Figure 4.3 Missing Values on Dataset

To fill the missing values, an analysis of the dataset is conducted to determine the appropriate values to be imputed. The strategy involves filling in the missing values for the *sub1* feature based on the *seg* feature and vice versa. The specific values chosen for imputation, based on this cross-feature analysis, are detailed in Tables 4.3 and 4.4. This method ensures a data-driven approach to handling missing values, thereby maintaining the integrity and usefulness of the dataset for further analysis.

Table 4.3 Filling *sub1* Missing Value

<i>seg</i> value	<i>sub1</i> filled with
account	account
financial	price

Table 4.4 Filling *seg* Missing Value

<i>sub1</i> value	<i>seg</i> filled with
account	financial
call failed	ring not answered
call rna	ring not answered
coverage	uncovered location
customer	account
internet problem	internet sering putus
price	more expensive
product	better product
whatsapp reminder	whatsapp no response

Once it is confirmed that there are no missing entries in the dataset, the next step involves conducting a conversion of data types for certain features. As illustrated in Figure 4.4, the features *date_activation* and *date_expired* are initially in an object data type but need to be converted to a date-time data type. This conversion is essential for accurate analysis and modeling, as it ensures that these features are in the correct format for any temporal analysis or operations that might be necessary in subsequent stages of the study.

```
df_candidate.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 73734 entries, 381 to 88581
Data columns (total 23 columns):
 #   Column              Non-Null Count  Dtype  
---  --
 0   customergender      73734 non-null  object 
 1   cstmr_age           73734 non-null  int64  
 2   tik_custid          73734 non-null  object 
 3   prepaidtype         73734 non-null  object 
 4   packageprice_idr    73734 non-null  int64  
 5   subscription_age    73734 non-null  float64 
 6   packagename         73734 non-null  object 
 7   date_activation     73734 non-null  object 
 8   date_expired        73734 non-null  object 
 9   cat1                73734 non-null  object 
10   sub1                73734 non-null  object 
11   seg                 73734 non-null  object 
12   closedate           73734 non-null  object 
13   profit_center_name  73734 non-null  object 
14   mapping_area        73734 non-null  object 
15   technology          73734 non-null  object 
16   sa_province         73734 non-null  object 
17   territory           73734 non-null  object 
18   city                73734 non-null  object 
19   district            73734 non-null  object 
20   sub_district        73734 non-null  object 
21   churn               73734 non-null  int64  
22   total_tickets       73734 non-null  int64  
dtypes: float64(1), int64(4), object(18)
memory usage: 13.5+ MB
```

Figure 4.4 Data Type Conversion

The subsequent step is to conduct data profiling, which aids in analysing and gaining a better understanding of the dataset. Data profiling involves examining both the numerical and categorical data present in the dataset.

Table 4.5 Numerical Data

	cstmr_age	packageprice_idr	subscription_age	churn	total_tickets
count	73734	73734	73734	73734	73734
mean	37	481560	2	1	1
std	19	378142	59	0	2
min	0	0	0	0	0
25%	26	375000	1	1	0
50%	33	425000	1	1	0
75%	44	425000	2	1	1
max	124	30000000	7987	1	49

Table 4.6 Categorical Data

	count	unique	top	freq
customergender	73734	2	male	65408
tik_custid	73734	73734	1100000215418	1
prepaidtype	73734	4	home	70294
packagename	73734	224	biznet home internet 1b rent	17622
cat1	73734	10	unable to retain	45053
sub1	73734	26	call failed	29045
seg	73734	59	ringing not answered	17549
closedate	73734	517	14/09/2023	479
profit_center_name	73734	206	branch denpasar	2725
mapping_area	73734	4	east	34432
technology	73734	4	gpon	73453
sa_province	73734	21	dki jakarta	17876
territory	73734	3	west	39300
city	73734	127	kota jakarta selatan	5712
district	73734	630	denpasar selatan	2059
sub_district	73734	2487	pluit	972

Based on Table 4.5, some potential issues have been identified, particularly concerning customer age:

- The minimum and maximum age for customer age of “0” and “124” might be an outlier and should be investigated.
- The maximum subscription age of “7987” might be an outlier or irrelevant data which might need further investigation.
- The minimum and maximum package price of “0” and “30000000” could be the real package price or the outlier and should need to investigate.
- For the target feature, churn, has a mean of 1, suggesting that there might be a class imbalance, and most customers churned.

In the dataset, for customer age, the analysis adheres to government policy which sets the minimum age for citizens to obtain an ID at 17 years old. Consequently, any customer age entries below 17 years are not in compliance with this policy and warrant

investigation. Additionally, an upper limit for customer age is set at 80 years old. This boundary is established to ensure data relevance and accuracy. Therefore, the investigation will extend to those entries in the dataset where customer ages are recorded as over 80 years, as these may be outliers or inaccuracies. This approach helps in maintaining the validity and reliability of the dataset, especially for analyses involving demographic factors like age.

Table 4.7 Customer with Age Under 17

	cstmr_age	packageprice_idr	subscription_age	churn	total_tickets
count	3805	3805	3805	3805	3805
mean	0	1040601	12	1	1
std	1	1211430	259	0	2
min	0	0	0	0	0
25%	0	325000	1	1	0
50%	0	575000	5	1	0
75%	0	1200000	6	1	1
max	16	30000000	7987	1	32

Table 4.8 Customer with Age Over 80

	cstmr_age	packageprice_idr	subscription_age	churn	total_tickets
count	1894	1894	1894	1894	1894
mean	123	863081	5	1	1
std	6	823747	1	0	2
min	81	0	0	0	0
25%	124	325000	4	1	0
50%	124	525000	5	1	0
75%	124	1150000	6	1	1
max	124	12500000	8	1	42

Based on table 4.7 and 4.8, it shown that there are 3805 and 1894 entries of data for customer under 17 years old and over 80 years old. After looking into the data, almost all customers that less than 17 years old are not really 17 years old, but it seems that the age data for these customers are missing. For customer who have an age over 80, it's impossible to have customers with ages of 124 years old. If these customers are simply removed from the dataset, a significant amount of information could potentially be lost. This loss could impact the comprehensiveness and representativeness of the dataset,

potentially affecting the accuracy of any analysis or conclusions drawn from the data. Therefore, careful consideration and alternative methods of handling these data points are required to preserve as much valuable information as possible while maintaining data integrity.

A significant number of individual customers have been noticed to fall below 17 years old or exceed 80 years old, with available age information lacking. Imputing these ages with statistical measures like the mean might introduce bias into the analysis. Additionally, considering the presence of business customers, associating an age attribute, typically linked to individuals, doesn't align logically as it pertains to human age, not company age. Thus, it has been decided to set aside the customer age column from the analysis.

Instead, the focus is being shifted to the subscription age as a more impactful factor influencing churn patterns. The duration of subscription or tenure with the service seems to hold more weight in understanding customer behaviour compared to the age of the customers themselves, especially given the presence of both individual and business clientele.

As the decision has been made to set aside the customer age, the focus can now be shifted to the subscription age. For the subscription age of “0” means that the customer has been subscribed for less than a year. However the subscription age of 7987 is outlier. The determination of the subscription age in the dataset, achieved through the computation of the difference between the *date_expired* and *date_activation* columns, presents a potential bias. This bias stems from instances where customers exhibit a churn status while their expiration date remains in the future. To mitigate this bias, an alternative approach is proposed: recalculating the subscription age or tenure create a new feature named *subscription_tenure*. This recalculation involves subtracting the close date from the date of activation for churned customers, while for active customers, it necessitates subtracting the current date from their activation date. The formula for this process are:

for churn customers:

$$\text{subscription age} = \text{close date} - \text{activation date}$$

for non-churn customers:

$$\text{subscription age} = \text{current date} - \text{activation date}$$

Table 4.9 Subscription Tenure

	cstmr_age	packageprice_idr	subscription_age	churn	total_tickets	subscription_tenure
count	73734	73734	73734	73734	73734	73734
mean	37	481560	2	1	1	2
std	19	378142	59	0	2	1
min	0	0	0	0	0	0
25%	26	375000	1	1	0	100%
50%	33	425000	1	1	0	100%
75%	44	425000	2	1	1	200%
max	124	30000000	7987	1	49	7

Now that a new feature has been introduced, the features that will not be used in the models and EDA can be dropped. The features to be dropped are *subscription_age*, *date_activation*, *date_expired*, and *closedate*.

Having identified instances of "0" and "30000000" package prices in the dataset, the next step involves undertaking more extensive monitoring and analysis of the package price.

```
[50] df_candidate[df_candidate['packageprice_idr']==0].shape
      (28, 19)

[55] df_candidate[df_candidate['packageprice_idr']==30000000].shape
      (1, 19)
```

Figure 4.5 Package Price Investigation

Based on figure 4.6, it shown that there are 28 entries for package price “0” and only 1 entry for package price “30000000”. Upon further examination, it is evident that all the customers that has a package price of “0” are associated with free accounts and complimentary accounts. As for the customer who has a package price of “30000000” is a postpaid customer which is not related with the dataset that is about prepaid customers. As all the data that contains package price “0” and “30000000” were not to related with the research, the entries that related with “0” and “30000000” in package price were eliminated. The result of this process shown in table 4.10.

Table 4.10 Package Price Investigation Result

	cstmr_age	packageprice_idr	churn	total_tickets	subscription_tenure
count	73705	73705	73705	73705	73705
mean	37	481343	1	1	2
std	19	362129	0	2	1
min	0	200000	0	0	0
25%	26	375000	1	0	1
50%	33	425000	1	0	1
75%	44	425000	1	1	2
max	124	12500000	1	49	7

Next, data integration will be conducted, and a new feature that can aid in processing the Exploratory Data Analysis (EDA) will be created. The 'seg' feature will be used as the basis for this integration and the creation of the new feature. This feature is chosen because it represents the result that the contact center receives about the outcome in the retention process.

The first column to be created in the data integration process will be the *call_status* feature. There will be two values in this feature: call failed and call success. To fill the values in the *call_status* feature, the basis will be the values from the *seg* feature. The condition for call failed are shown in table 4.11. As for call success, it will use the rest of values from *seg* feature.

Table 4.11 Call Failed Value

<i>call_status</i> Value	<i>seg</i> Conditions
call failed	ringing not answered
	whatsapp no response
	wrong number
	busy number
	number out of service
	empty place
	hard to contact
	call rna
	lokasi kosong
	ring not answered

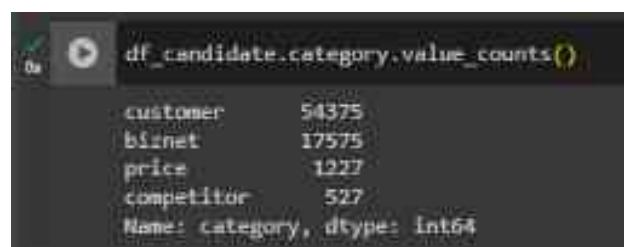


```
[61] df_candidate.call_status.value_counts()

call failed      46218
call success     27487
Name: call_status, dtype: int64
```

Figure 4.6 *call_status* Feature Values

Next, a new feature named *category* will be created to provide information about whether customers churn or not, based on the *seg* feature. The *category* feature will be simplified, like the *cat1* feature, as it will consist of only 4 values. The process of creating the *category* feature will follow the same approach as when creating the *call_status* feature, as shown in figure 4.8. After this process, *cat1* and *seg* features can be dropped to simplify the dataset.



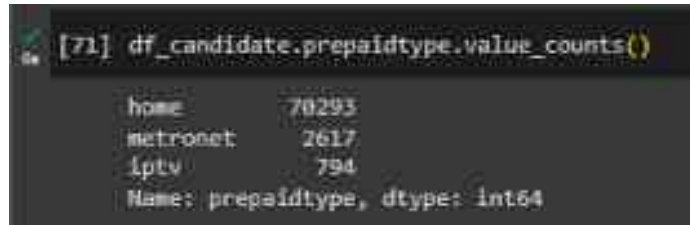
```
df_candidate.category.value_counts()

customer      54375
biznet        17575
price          1227
competitor      527
Name: category, dtype: int64
```

Figure 4.7 *category* Feature Values

The last feature to be created is *cust_seg*. This feature will categorize the dataset into two values: 'business' and 'individual.' To determine the *cust_seg* value, the *prepaidtype*

feature will be used as the basis. In the *prepaidtype*, there are three values that can represent business or individual customers. 'home' and 'iptv' will represent individual customers, while 'metronet' will represent business customers.



```
[71] df_candidate.prepaidtype.value_counts()

home      70293
metronet   2617
iptv       794
Name: prepaidtype, dtype: int64
```

Figure 4.8 prepaidtype Values



```
df_candidate.cust_seg.value_counts()

individual  71887
business    2617
Name: cust_seg, dtype: int64
```

Figure 4.9 cust_seg Values Result

Now that the dataset has been cleaned, consisting of 73,704 entries with 20 columns, the next step is to proceed with Exploratory Data Analysis (EDA) for the visualization process.

4.3 Exploratory Data Analysis (EDA)

Now that the data has been cleaned, the next step is to conduct a more in-depth analysis using visualization techniques or Exploratory Data Analysis (EDA). By doing EDA, it helps identify and visualize factors contributing to customer churn, decide which factors to include or exclude in the model, assists in feature selection, and help to identify and address class imbalance which is essential for building effective predictive models. EDA is also a powerful tool in providing valuable insights in marketing that can inform strategy and decision-making by visualizing the data. The visualization process will be carried out one by one based on the categories specified in table 4.12.

Table 4.12 Feature Dataset

Category	Feature
Demographic Information	customergender, cstmr_age
Subscription Details	packageprice_idr, subscription_tenure, packagename
Billing and Payment Information	profit_center_name
Communication and Service Interaction	seg, total_tickets, call_status, category
Geographic Information	sa_province, territory, city, district, sub_district
Technology and Network	technology
Customer Account Information	tik_custid, prepaidtype, cust_seg
Target Variable	churn

4.3.1 Demographic Information.

In this phase, the analysis will focus on examining the relationships between the features and the target variable, as well as gaining marketing insights from the results.

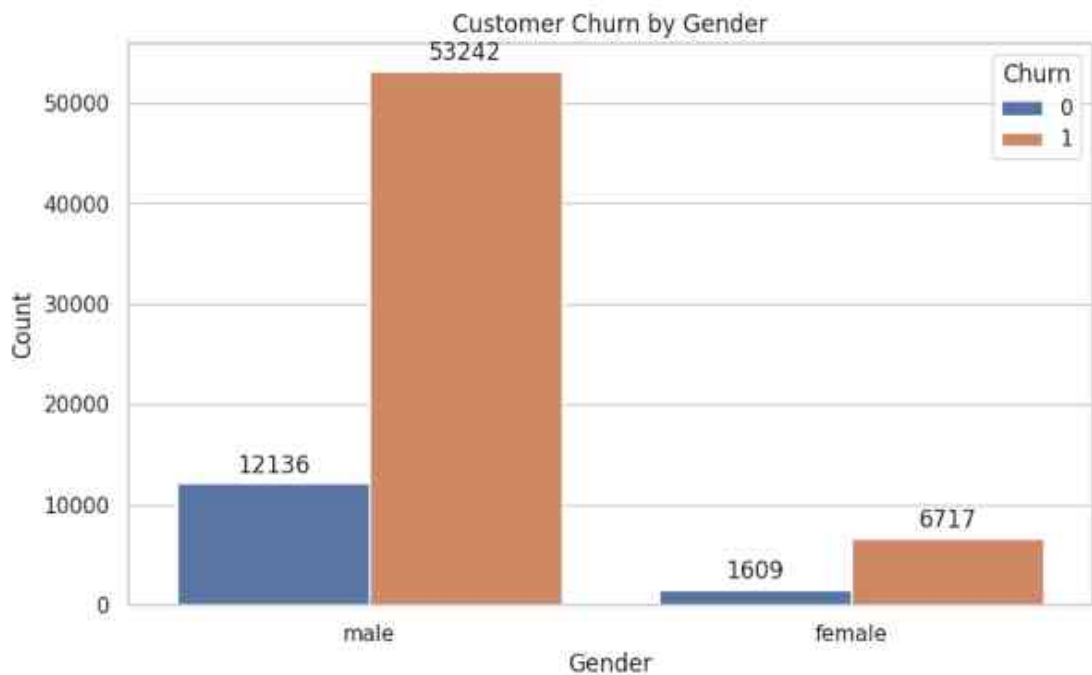


Figure 4.10 Customer Churn by Gender

The associated count plot shows the association between customer gender and churn. For each gender group, the bars reflect the number of customers who have churned (1) and those who have not churned (0).

Furthermore, the churn rates computed are as follows:

- Female customers have a churn rate of approximately 80.7%
- Male customers have a churn rate of approximately 81.4%.

This implies that male consumers have somewhat greater turnover, although the difference is not significant. To ensure the relationship between customer gender and churn, a statistical test will be conducted. A chi-squared test will be conducted to assess the statistical significance of the relationship between customer gender and churn. Additionally, to evaluate the strength of this relationship, Cramér's V will be computed. These analyses will help determine if customer gender is predictive enough to be included in a churn prediction model. The results of the chi-squared test and Cramér's V are as follows:

- Chi-squared Statistic (Chi2): 2.778
- p-value: 0.0956
- Degrees of Freedom: 1
- Cramér's V: 0.0049

The chi-squared test generates a p-value of around 0.096, which is somewhat higher than the standard alpha threshold of 0.05 used to establish statistical significance. This shows that the relationship between customer gender and turnover is not statistically significant. Cramér's V is a measure of association strength that goes from 0 to 1, with 0 indicating no connection and 1 representing perfect linkage. Cramér's V is roughly 0.0049 in this case, indicating an extremely weak link.

Based on these findings:

- Predictive Modeling: due to the lack of statistical significance and the extremely weak correlation with turnover, customer gender may not be a significant predictor of churn on its own.

- **Marketing Insights:** because the influence of gender on turnover seems to be small, marketing initiatives targeting churn may not need to be significantly gender targeted.

This does not imply that gender should be fully ignored in marketing initiatives. Gender may still influence elements of consumer behavior and preferences that are unrelated to churn. Different genders, for example, may react differently to marketing campaigns, product features, or customer service approaches, which may still be useful information for a marketing team.

Now, let's proceed to analyze customer age. The variable *cstmr_age* in the dataset has a wide range of values, including probable outliers such as ages of 0 and an extremely high age of 124 years. These outliers might be data entry mistakes or unusual circumstances that are not reflective of the average consumer base. To overcome this problem and assure more accurate and representative analysis, the EDA's method entails imposing constraints on consumer age data. A restriction on the Exploratory Data Analysis (EDA) will be imposed, based on previous actions in data preparation, excluding customers under the age of 17 and above the age of 80. This choice was made to reduce the influence of any outliers and data inconsistencies on the research.

Because of these inconsistencies, *cstmr_age* was not included as a predictor in the churn prediction model. Outliers and a skewed distribution of age data may cause mistakes in the predictive modeling process. As a result, this decision restricts the investigation of the possible influence of age on customer churn. However, *cstmr_age* was retained for marketing analysis and insights, where its distribution throughout the client base within the stated age range might give useful information for focused marketing initiatives. Future research might benefit from a more refined dataset in which age data is reliably collected and outliers are adequately controlled, allowing for a thorough examination of the age element in customer churn prediction.

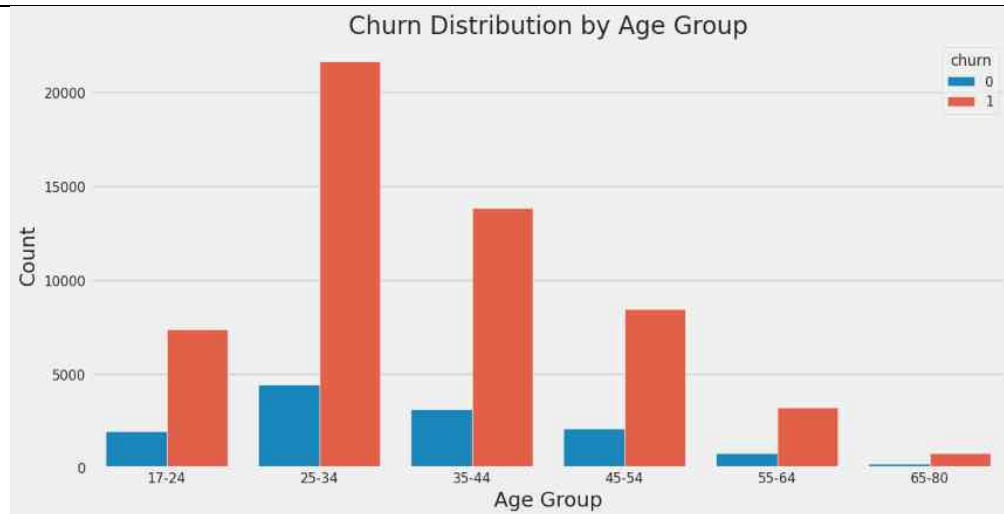


Figure 4.11 Churn Distribution by Age Group

Figure 4.12 shows how customer turnover varies across different age groups. This plot count visualization clearly describes patterns that can be identified in churn levels in different age groups, highlighting variability and trends within each group. Especially, increased churn rates were observed in the age group of 25-34. This increase indicates certain vulnerabilities or higher levels of dissatisfaction that are prevalent among customers in this demographic segment. Such trends may indicate a need or unmet expectation that is more pronounced at this young age group.

As the older age group is advanced too, a gradual decrease in churn rates is revealed by the data. Although this decline is not sharp, it indicates a slightly higher retention rate among older customers. However, it is important to note that the differences in unemployment rates between these older groups are insignificant, indicating the level of balance in customer satisfaction or loyalty across this age category.

This information can be used to tailor marketing and retention strategies for specific age groups, especially focusing on those with higher growth rates. Understanding the reasons behind higher descent in a particular age group can inform targeted interventions. For example, increasing age in the age group 25-34 requires investigating unique factors that drive this trend, which can lead to more effective and age-specific strategies aimed at increasing customer engagement and reducing age.

Overall, Figure 4.12 not only emphasizes the importance of considering age as a factor in a customer retention strategy but also points to the need for a more nuanced understanding of customer needs and preferences in different age groups. This knowledge is essential to adapt marketing strategies and increase customer engagement efforts, especially for the age group of 25-34, where churn is much higher.

4.3.2 Subscription Details

For subscription details category, let's start with the package price. A boxplot visualization will be used to determine if there is any correlation between package price and churn.

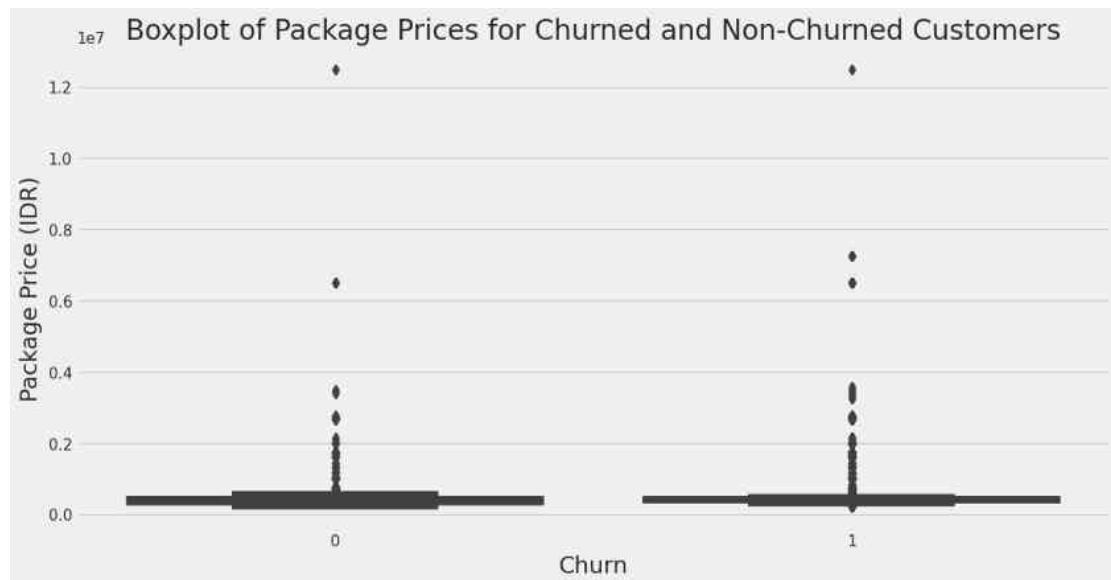


Figure 4.12 Boxplot of Package Price and Churn

From the quartile analysis of package prices for churned and non-churned customers it shown that:

Churned Customers:

- 25th Percentile (Q1): 375,000 IDR
- 50th Percentile (Median, Q2): 425,000 IDR
- 75th Percentile (Q3): 425,000 IDR

Non-Churned Customers:

- 25th Percentile (Q1): 325,000 IDR
- 50th Percentile (Median, Q2): 425,000 IDR
- 75th Percentile (Q3): 425,000 IDR

Based on these findings, it can be observed that:

- **Similar Median Prices:** The median package price (50th percentile) for both churned and non-churned customers is the same at 425,000 IDR. This suggests that, on average, both groups have similar package prices.
- **Price Range Variance:** While the median prices are identical, there's a difference in the lower quartile (25th percentile) between churned (375,000 IDR) and non-churned (325,000 IDR) customers. Churned customers tend to have a higher price starting point.
- **Limited Price Variability:** Both groups have identical values for the upper quartile (75th percentile) at 425,000 IDR, indicating that most customers, whether churned or non-churned, fall within this price range.

The implications for these observations are:

- **Higher Starting Price for Churned Customers.**
According to the data investigation, churned consumers are more likely to begin their subscriptions with higher-priced packages than their non-churned counterparts. The higher 25th percentile (lower quartile) stated for churned customers supports this tendency. This finding implies a possible correlation between greater initial package cost and turnover. This might point to a situation in which consumers who choose more costly packages join the service with higher expectations or a greater sensitivity to the perceived value they get. As a result, if the service falls short of their expectations, these clients may be more likely to leave it. The connection between higher initial pricing and future churn emphasizes the significance of matching pricing strategies with consumer expectations and value delivery, particularly in higher price ranges.
- **Uniformity in Median and Upper Quartile.**

The examination of package pricing within the dataset demonstrates significant equality in the middle and higher ranges of package costs across both churned and non-churned clients. This is especially noticeable since the median (50th percentile) and upper quartile (75th percentile) package costs are the same across these categories. Such price consistency might be suggestive of the company's consistent pricing policy, suggesting a method that applies evenly all over the whole consumer base. Alternatively, this might indicate a concentration of clients, regardless of churn status, inside comparable package options. This might suggest that most consumers, whether they churn or not, choose packages within a certain price range. This price range is most likely the company's most popular or typical pricing tier. The existence of a popular price tier that appeals widely to the customer base, as indicated by the pricing similarity at these center and higher percentiles, suggests a potentially successful pricing approach that responds to current client preferences.

Several marketing insights that can be derived from the data include:

- Pricing tactics that are specific to the customers or that provide additional value within higher-priced packages may be critical to maintaining customers. Promotions or expanded service features may help to justify higher rates.
- It is critical, particularly for premium packages, to ensure that the value offer corresponds with client expectations. This might include better customer communication, greater service quality, or extra benefits available only to higher levels.
- It might be advantageous to identify and capitalize on the most popular pricing ranges. A particular focus on certain categories via targeted marketing campaigns or bundle offers might help to acquire a larger consumer base.
- It might be beneficial to have a tiered pricing plan that accommodates various client demands and preferences. Offering starting prices, loyalty discounts, or adjustable package choices are examples of this.

Based on the analysis, the feature *packageprice_idr* will be use as a feature for the prediction model.

Let's now move on into *packagename* to find the relation between this feature with churn.

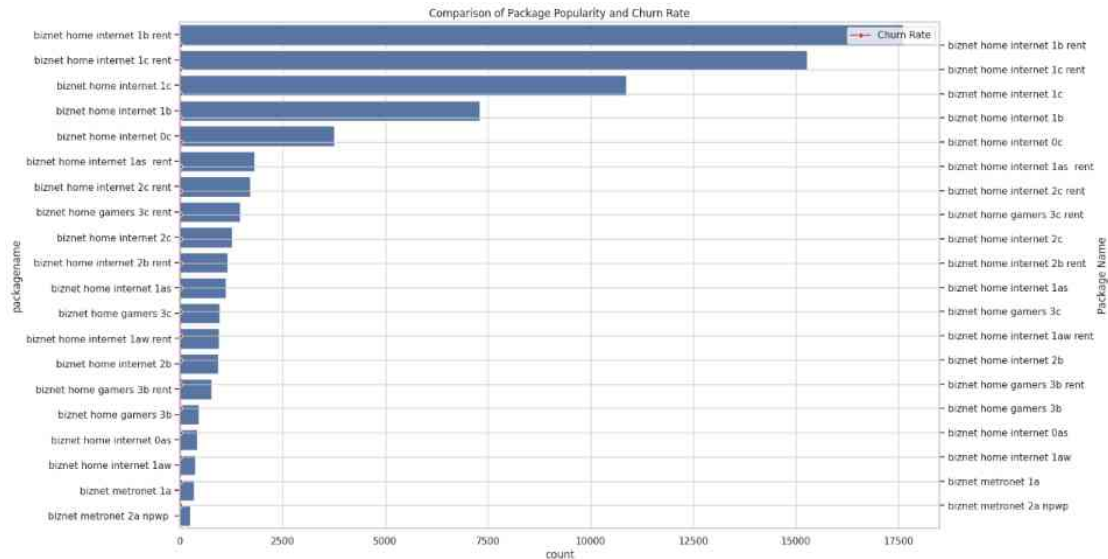


Figure 4.13 Top 20 Popular Package Name

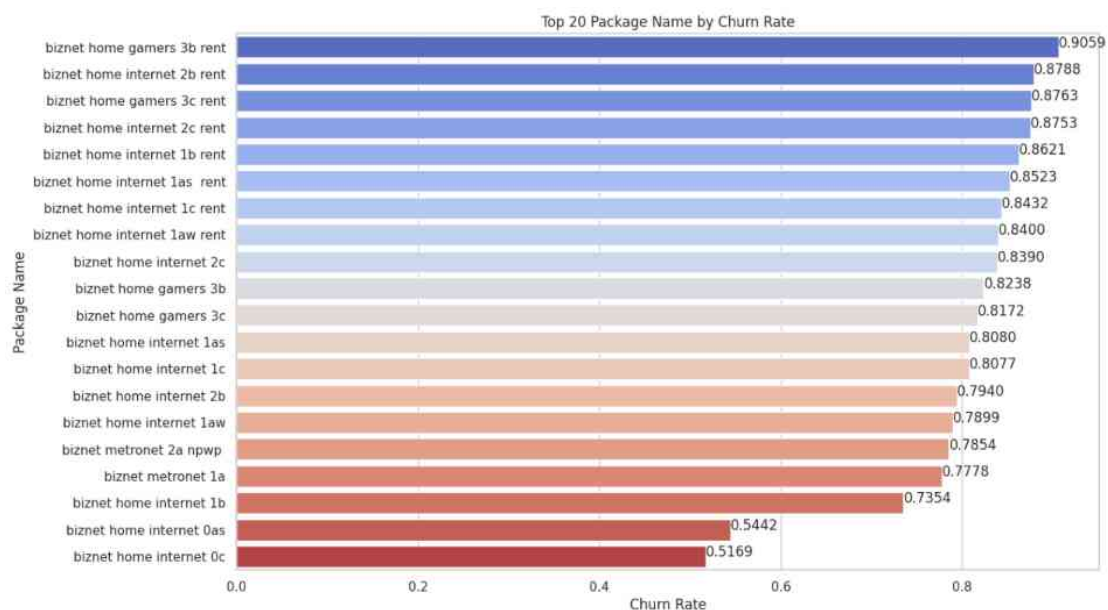


Figure 4.14 Top 20 Package Name by Churn Rate

The top three most popular packages are "biznet home internet 1b rent," "biznet home internet 1c rent," and "biznet home internet 1c," according to the data research. Notably, the popularity of a package does not always correspond with the churn rate. More

contrary, several of the most popular packages have lower churn rates, but unexpectedly, others have higher rates. This occurrence suggests that factors more than just popularity affect churn.

Furthermore, it was observed that some products with a lower subscriber base had greater churn rates. This trend may indicate dissatisfaction on the part for customers or an inconsistency between what these products really deliver and what customers really want or anticipate. These findings highlight how crucial it is to dig more into package performance and consumer preferences to improve the products and better match customer needs.

On the other hand, the data analysis also provides crucial insights for marketing. Investigating the reasons behind the high churn rates of specific products that are not as popular with customers is crucial. There are several reasons for this tendency, such as poor pricing tactics, a failure to satisfy customers, or a lack of alignment with market needs. To tackle these problems, it is a good idea to actively seek out and carefully consider customer feedback, as this may help identify areas that might be improved upon as well as the underlying reasons why customers are not satisfied.

Additionally, packages with low churn rates and wide distribution provide a great chance for marketing campaigns. These packages seem to be well-received by customers, since they have shown their capacity to attract and retain customers. One tactic to think about for packages with high churn rates is combining them with more popular services or redesigning them to make them more appealing and reduce customer churn.

It would also be wise to create and put into action custom retention plans for the packages that are most impacted by churn. Fostering client loyalty and decreasing the chance of churn might be greatly aided by tailored rewards, alluring loyalty programs, and improved customer assistance. In addition to raising client happiness, these focused initiatives may strengthen the services' entire value proposition.

To confirm the relationship between package name and churn, a statistical test will be conducted, and the results will be analyzed:

Chi-square result: 3207.5579730416357

p-values: 0.0

Based on the statistical result, there is a significant association between the package chosen by a customer and their possibility to churn. This is in line with the previous marketing insight that marketing strategies specialize in a specific package could be effective in managing churn. In line with the previously analyzed package price, both features will be considered as candidates in the creation of a predictive model.

Next process is to analyze about subscription tenure in the dataset. Figure 4.16 indicates that the subscription tenure research of the customer base shows a trend towards shorter tenures, with a peak observed between 0.8 and 1.3 years. This distribution, which ranges from 0 to 7 years, shows a wide range of client involvement and loyalty. During the early phases of the customer lifecycle, retention initiatives including improved welcoming experiences, early engagement programs, and attractive introductory offers should be prioritized to increase retention. Customers who have been with the company for up to seven years are a very loyal group that might benefit from customized loyalty programs and special incentives.

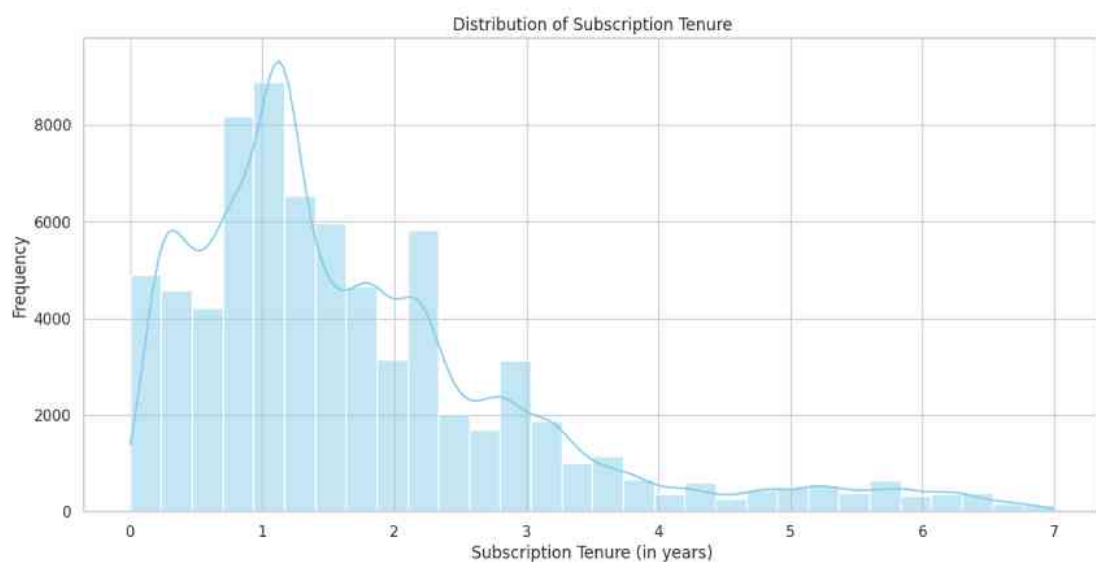


Figure 4.15 Distribution of Subscription Tenure

Given the huge number of customers that leave within short periods of time, an in-depth investigation of the main reasons of early churn is required. Surveys and exit interviews may provide further insight into the reasons behind exits and contribute to developing retention tactics. Customized offers and information, together with segmented marketing methods, may prove to be very successful. Using data analytics to pinpoint the crucial points in a customer's stay when they are most likely to leave might help with the creation of focused marketing campaigns and exclusive deals.

4.3.3 Billing and Payment Information

In the next phase of the analysis, the focus will be on the category of billing and payment information. The dataset currently provides this category with the attribute *profit_center_name*, which is the customer's location as determined by the company's branch network. It's crucial to remember that, in this situation, the dataset has limitations. In particular, it lacks specific elements like payment method, billing information detailed, and payment frequency that would provide deeper insights into billing and payment dynamics. The ability to provide an in-depth examination of customers' payment patterns and preferences may be impacted by the lack of detailed data on payment methods and cycles.

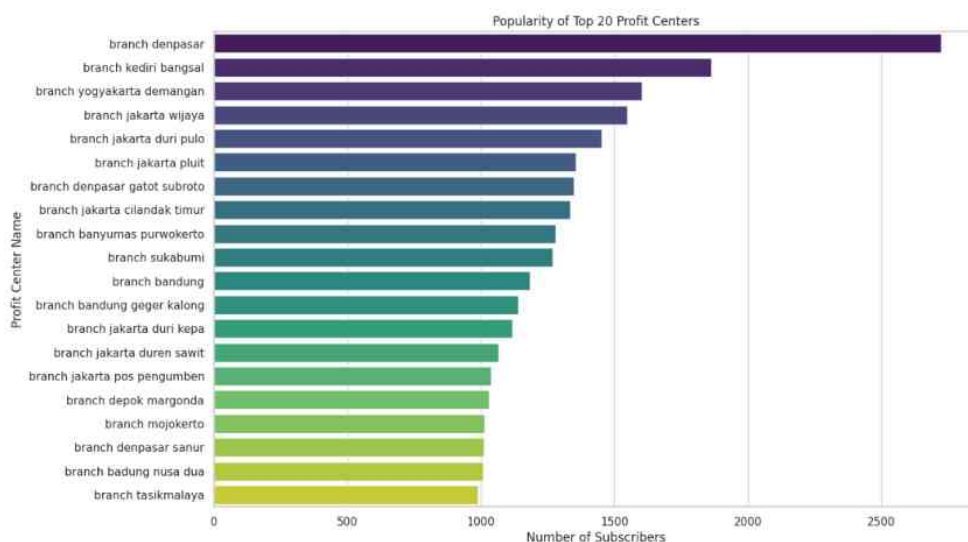


Figure 4.16 Top 20 Profit Centers Name

Based on the visualization, some profit centers have greater rates of customer churn than others, which makes them important targets for client retention initiatives. With 206 unique entries in the *profit_center_name* category, an in-depth examination of every location would take a lot of time and would not provide meaningful insights. More crucially, the predictive model runs the risk of overfitting and unnecessary complexity when it is focused too finely. As a result, a different strategy has been adopted. Instead of investigating every profit center separately, other variables contained in the dataset will be used to represent more comprehensive trends and patterns. This strategy will enable us to preserve a reasonable level of model complexity while capturing the crucial marketing insights, guaranteeing an enhanced and more general investigation.

4.3.4 Communication and Service Interaction

The focus now shifts to the area of communication and service interaction. Three crucial factors will be the focus of the investigation in this phase: *total_tickets*, *call_status*, and *category*. It should be noted that since the *seg* variable has already been transformed into new variables during the data preparation phase, it will not be discussed further. The focus will be on examining three key metrics: *total_tickets* (the total number of service tickets a customer has raised); *call_status* (the result of service calls); and *category* (the major category of customer retention call result). To understand how interactions with service and communication affect customer retention, the connections between these factors and customer churn will be investigated.

The first factor to be visualized is *total_tickets*, representing the number of tickets related to service disruptions that have been made by the customer.

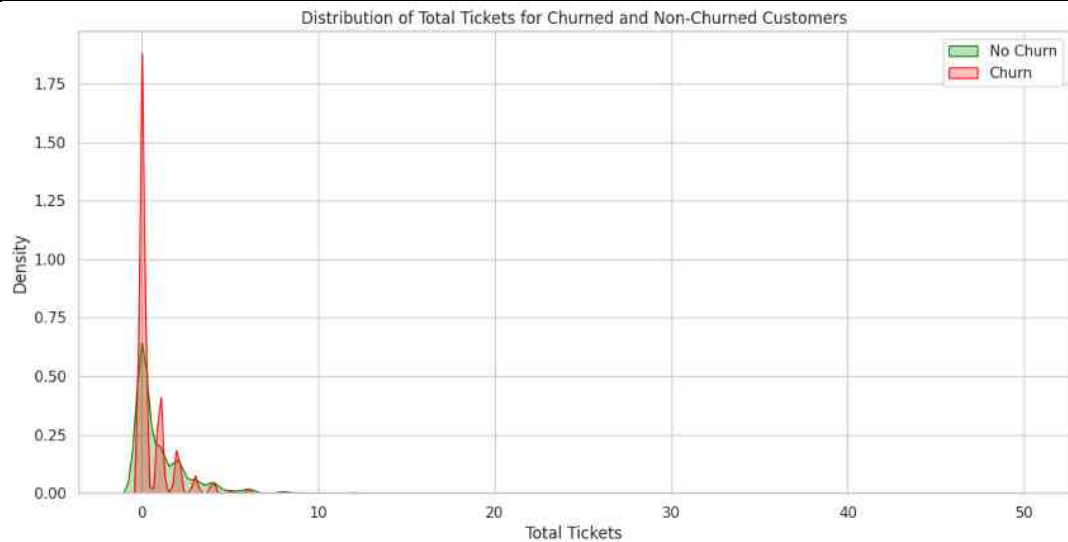


Figure 4.17 Distribution of Total Tickets for Churned and Non-Churned Customers

The statistical test:

Chi-Square Statistic: 1759.24

p-value: 0.0

Degrees of Freedom: 31

According to the *total_tickets* investigation, clients who have no tickets are more likely to churn, maybe because of poor involvement or ignorance of available help options. On the other hand, there is no clear correlation between a greater churn rate and the quantity of tickets raised, indicating that churn is more heavily impacted by customer happiness and service quality. Customer loyalty may be increased via the use of loyalty programs or unique rewards.

In addition, the Chi-Square test rejects the first visual analysis and finds a substantial correlation between total tickets and turnover. This demonstrates how crucial statistical and visual analysis are to comprehending nuanced consumer behavior. The test's results highlight the slight, although not immediately obvious visually, impact total tickets play in customer churn and how important it is for predictive modeling. Developing client retention tactics that work better is made possible by this comprehensive analytical technique.

Next let's head into the call status results from customer retention conducted by contact center.

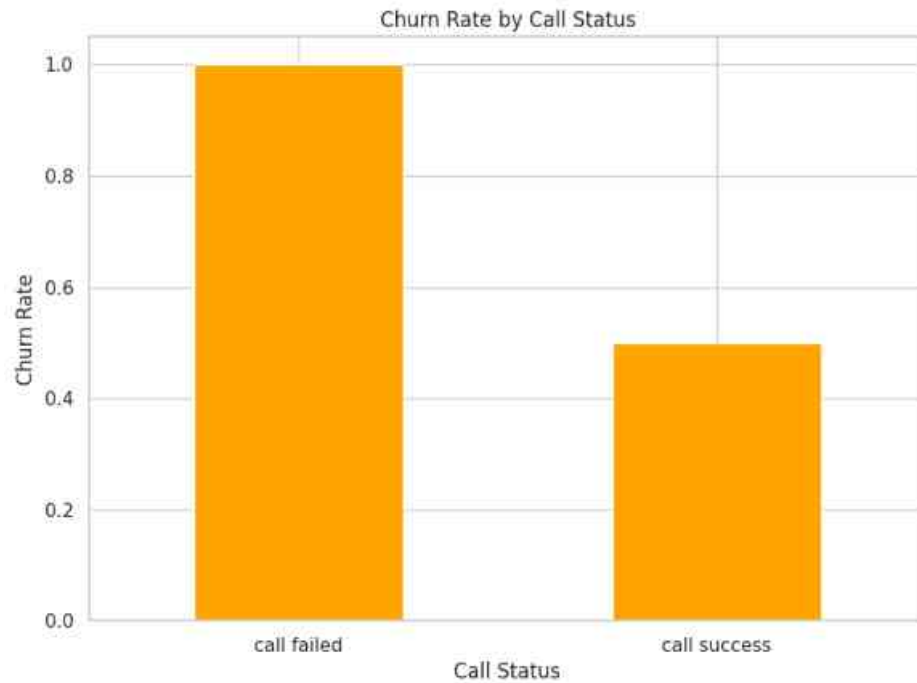


Figure 4.18 Churn Rate by Call Status

The statistical test:

Chi-Square Statistic: 28400.75

p-value: 0.0

Degrees of Freedom: 1

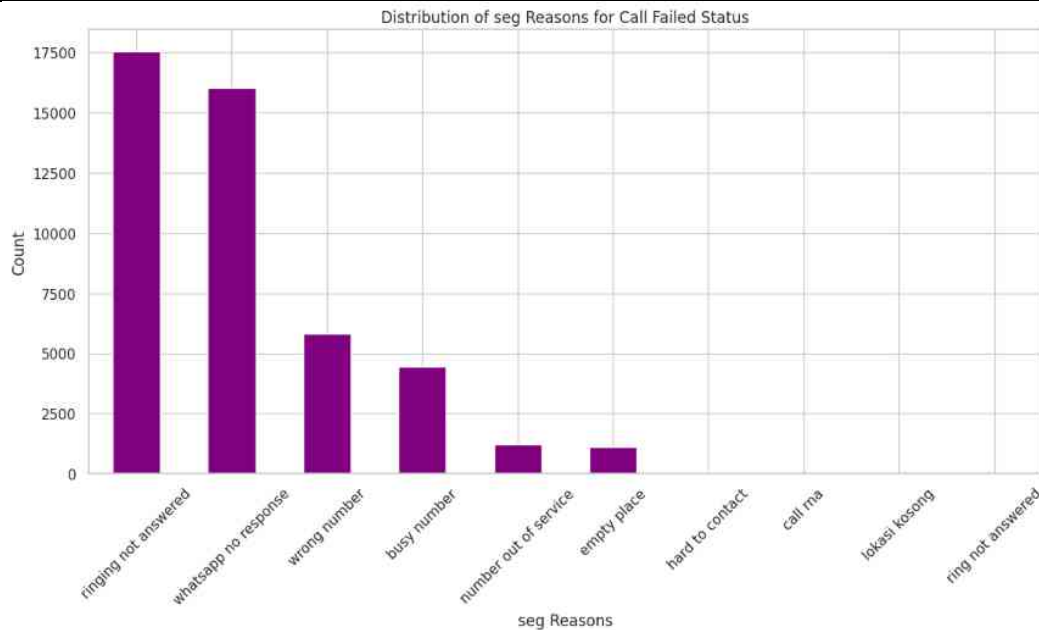


Figure 4.19 Distribution of *seg* Reason for Call Failed Status

The data shows a notable difference in churn rates across various call statuses, with unsuccessful calls having a much higher churn rate. The outcome of service calls—whether they were successful or unsuccessful—seems to be a key factor in forecasting customer turnover, according to the statistical data. The customer appears to be the main cause of call failures. This covers problems like calls coming in at odd hours, caller IDs becoming unidentified, or customers being cautious to take calls from unfamiliar numbers. These elements point to possible areas where communication with customers strategies should be strengthened.

Maintaining up-to-date and correct client contact information is essential for improving call success rates, minimizing unsuccessful connection attempts, and guaranteeing that the business has up-to-date information.

Key initiatives also include strengthening customer connections, increasing interaction via various channels such as email, text messaging, social media, or app alerts, and shifting to a more proactive customer support strategy. These techniques have the potential to decrease churn, boost recognition and trust, and eventually improve total customer happiness by addressing consumer hesitancy.

In conclusion, call success rates may be greatly increased and customer retention can be substantially impacted by a multidimensional strategy that incorporates improved data management, customized communication tactics, and proactive customer service.

Let's now review the summary of the factors that led to the outcome of the customer retention call.

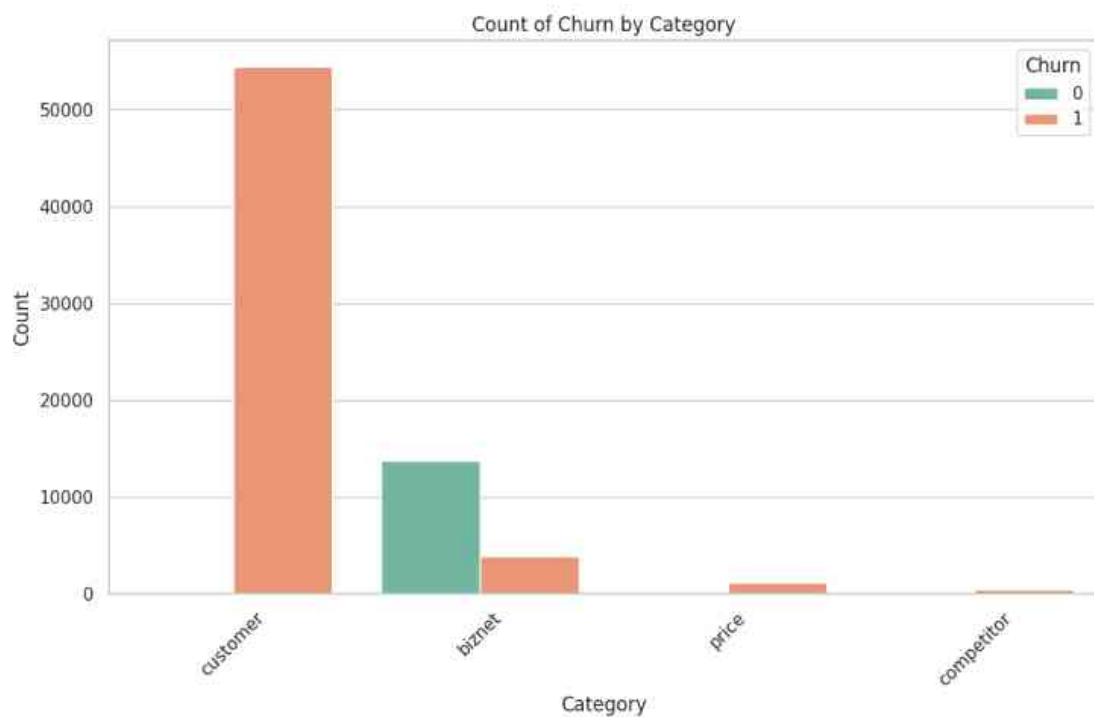


Figure 4.20 Churn by Category

The statistical test:

Chi-Square Statistic: 53692.48

p-value: 0.0

Degrees of Freedom: 3

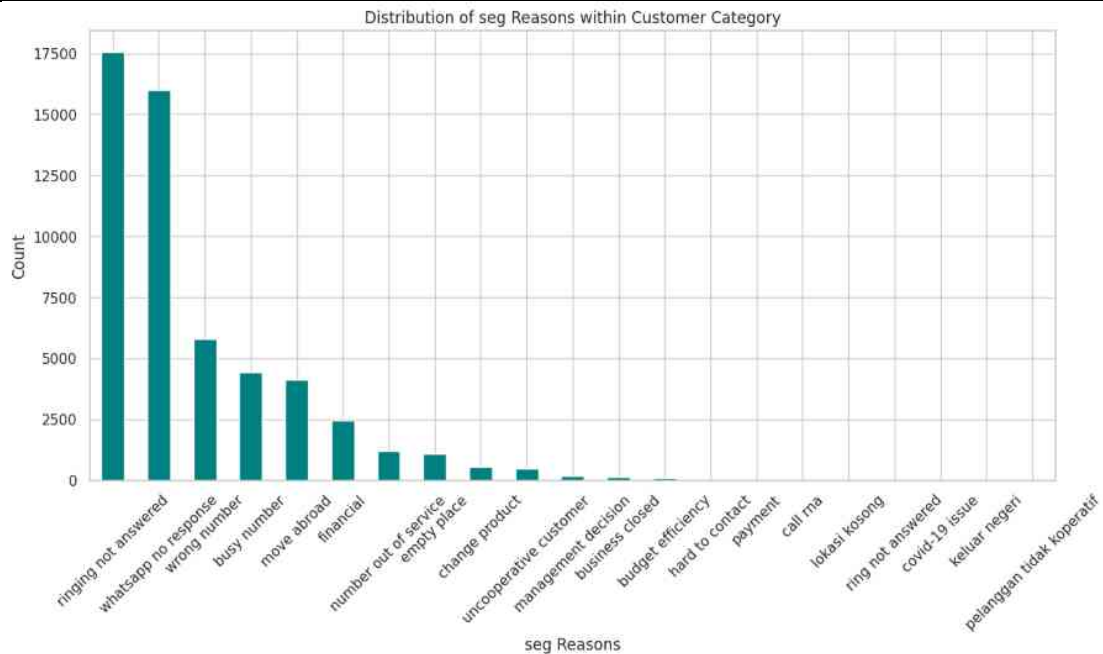


Figure 4.21 Distribution of *seg* Reasons Within "customer" Category

The visualization leads us to the conclusion that the *category* feature and *call_status* have similar underlying problems, which mostly include difficulties reaching out to consumers. The marketing tactics obtained from the research of *call_status* may be applied to *category* as well, emphasizing the significance of efficient communication and client involvement.

Similar to *call_status*, it is clear from the Chi-Square test findings that *category* has a substantial correlation with churn. This suggests that these characteristics play a crucial role in understanding and predicting customer churn. A common pattern seen in both *call_status* and *category* are communication issues. This highlights the need for better contact tactics, accurate and current customer information, and improved overall quality of customer care.

As a result, *category* — like *call_status* — can be a useful characteristic in churn prediction models. To increase customer retention, it offers insights into certain issues with customers or behaviors that may be resolved with customized marketing and customer care plans.

4.3.5 Geographic Information

In the geographic information analysis, the relation between the features *sa_province* and *territory*, which represent the general geographic information of the customers, will be analyzed. For *city* features, the analysis will be utilized to find the relation between churn and the distinction of urban/rural areas. Urban will be represented by “kota” and rural will be represented by “kabupaten” from *city* entries. As for *district* and *sub_district* features, the analysis of the relationship between churn and urban/rural areas will be conducted only if there is a need to delve deeper into the geographic information of customers.

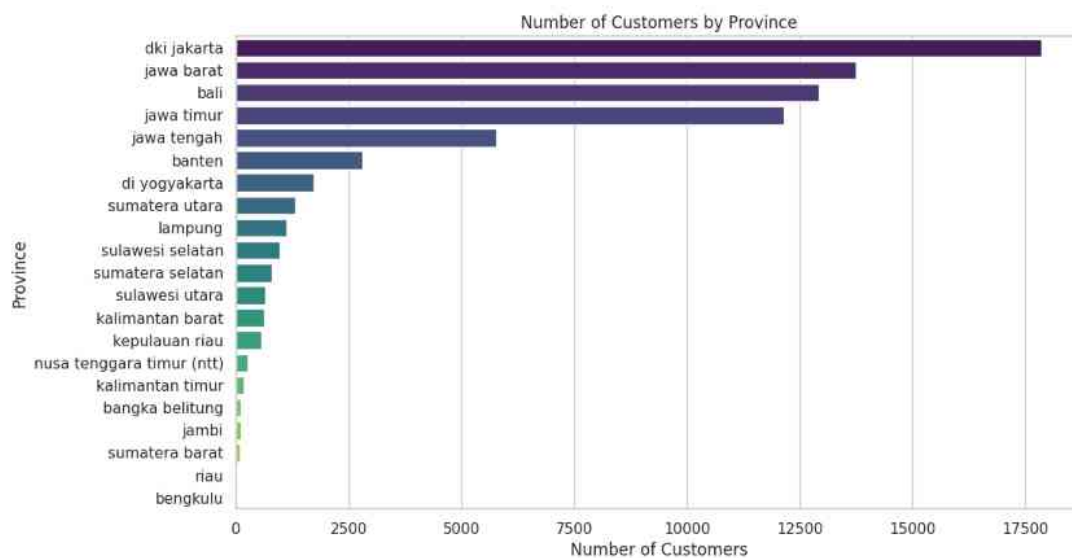


Figure 4.22 Number of Customers by Province

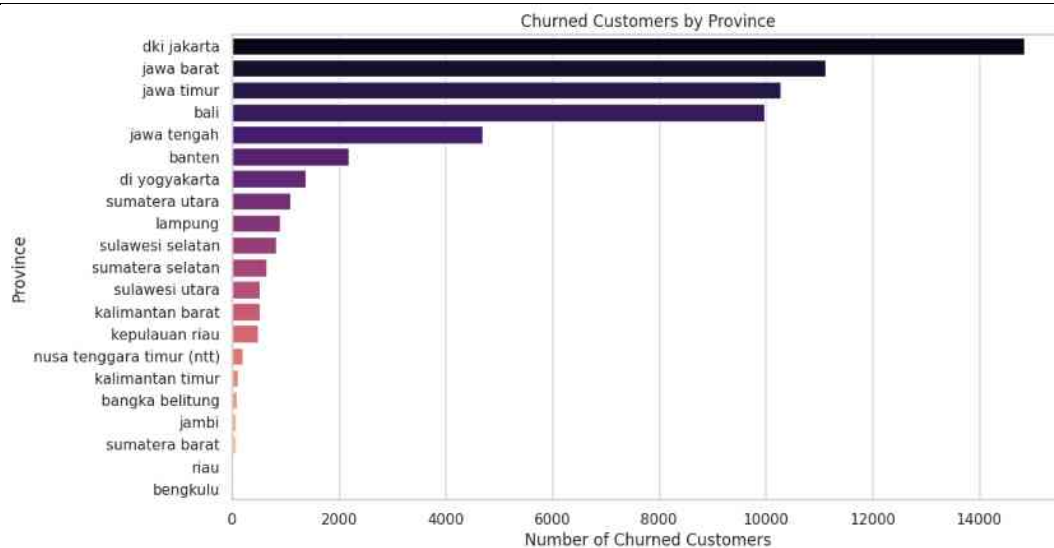


Figure 4.23 Churned Customers by Province

The dataset analysis indicates that DKI Jakarta has the greatest number of customer churn incidents in addition to having the biggest customer base. This situation indicates a higher number of churn instances is a reasonable result of DKI Jakarta's position as a significant province with a significant base of customers. The high rate of customer turnover may be attributed to several factors, such as intense competition in the market, unique consumer preferences, or certain service-related issues that are common in the area. This emphasizes the need for DKI Jakarta to implement specialized retention efforts. These tactics might include improving the quality of client service, creating customized marketing campaigns, or providing attractive service bundles. Furthermore, doing comparison surveys between DKI Jakarta and provinces with lower rates of customer churn might provide important information that would aid in the development of more successful customer retention strategies.

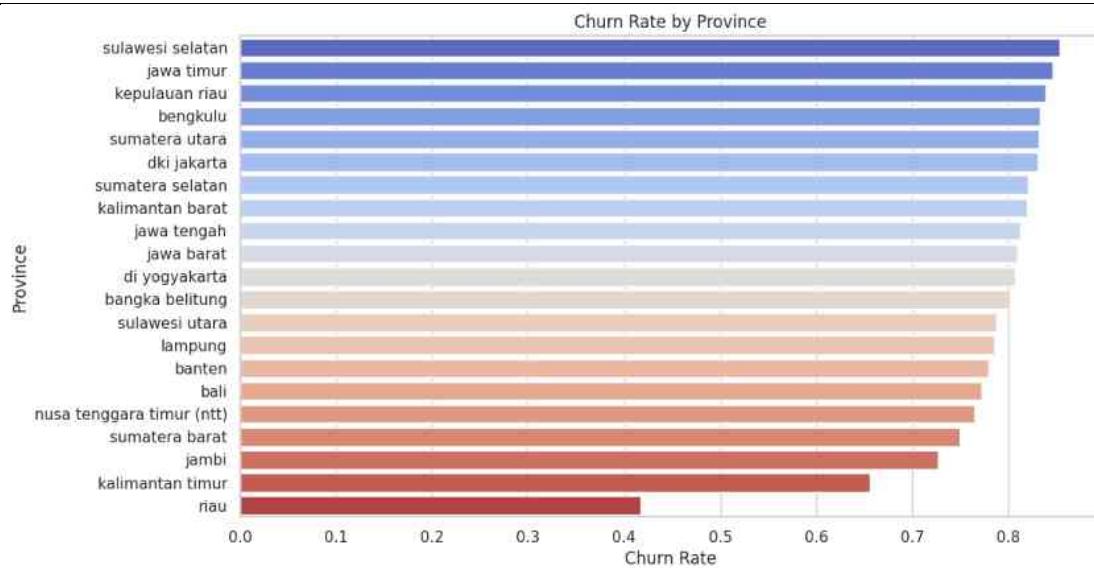


Figure 4.24 Churn Rate by Province

While DKI Jakarta leads in terms of the total number of customer churns, the analysis reveals that provinces such as Sulawesi Selatan, Jawa Timur, and Kepulauan Riau shows some of the highest churn rates. This insight highlights the need of creating marketing plans customized to various regions. It is essential to understand and deal with the unique customer needs, market dynamics, and service-related issues in these high-churn areas. It is probably more efficient to customize goods and services to each province's unique needs than to implement a standard strategy across all areas. Furthermore, it is critical to keep an eye on how these methods work and to be prepared to adapt as needed in reaction to changing market circumstances and the efficacy of the policies put in place.

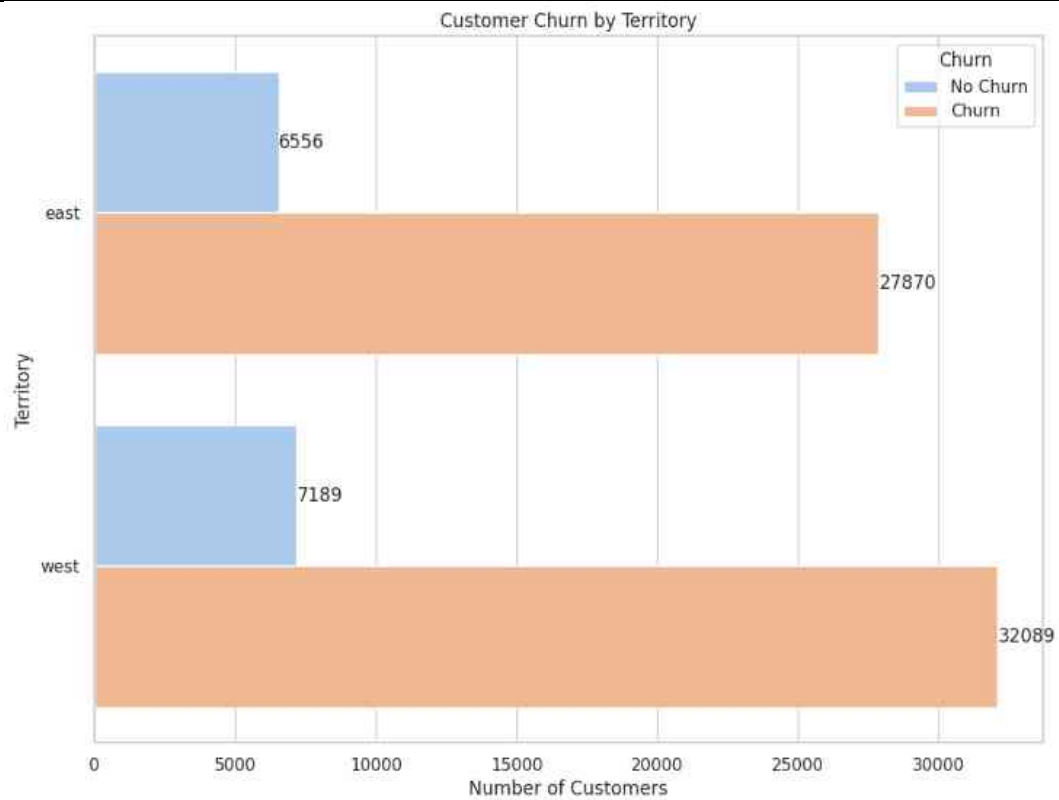


Figure 4.25 Customer Churn by Territory

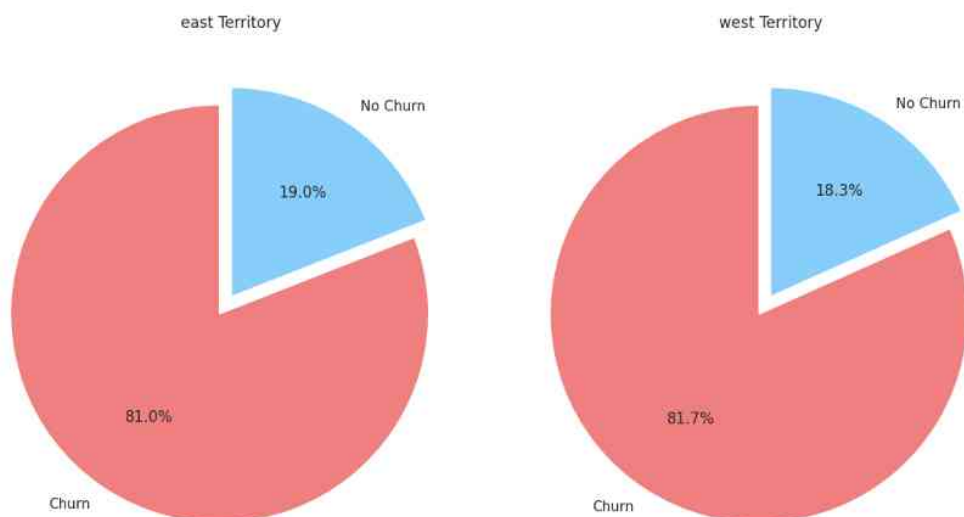


Figure 4.26 Percentage Customer Churn by Territory

The investigation of the *territory* feature revealed unique trends in the distribution of customer churn throughout the territories. Particularly, the statistics for both churned

and non-churned customers are greatest in the West area. This finding aligns with the previous analysis of the *sa_province* feature, which showed that the province with the largest client base was DKI Jakarta, which is situated in the West territory. Territories with high churn rates might benefit greatly from the same marketing tactics that have been recommended for provinces with high churn rates. These tactics include individualized customer engagement, improving service quality, competitive analysis, and developing customer loyalty. On the other hand, in regions where customer turnover is minimal, a comprehensive examination may reveal efficient methods that improve customer retention. Especially in high-risk areas, these data help direct customized strategies to improve consumer happiness and loyalty.

Further analysis will look more deeply at the influence of churn on urban and rural areas.

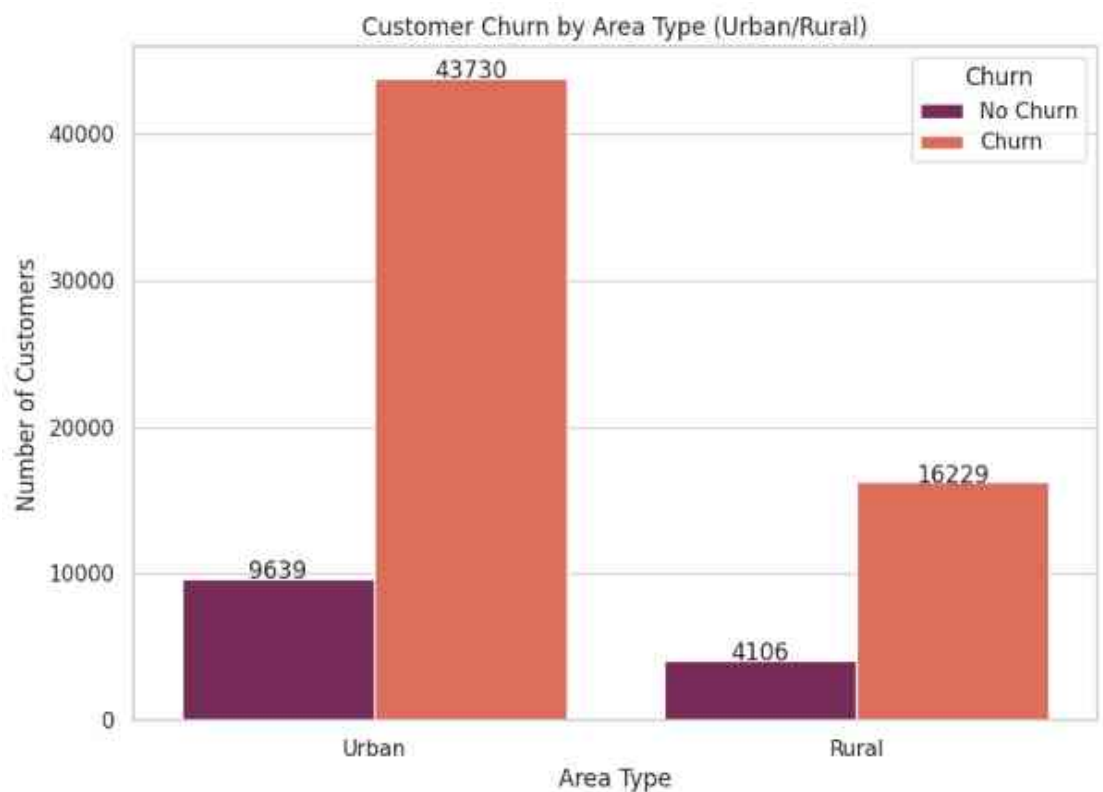


Figure 4.27 Customer Churn by Area Type (Urban/Rural)

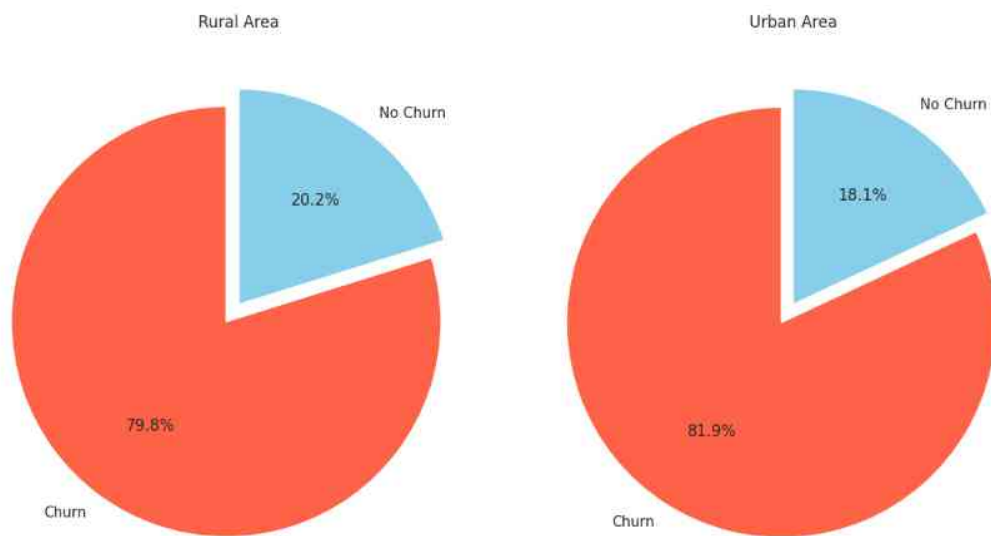


Figure 4.28 Customer Churn Percentage by Area Type (Urban/Rural)

Churn rate analysis shows that urban customers are more likely than those living in rural areas to discontinue services with the internet service provider, as seen by the higher churn rate. In particular, the non-churn rate for customers in rural areas is 20.2%, which is somewhat higher than the 18.1% rate for urban customers. This small but significant difference shows that rural clients typically have higher retention rates. One reasonable explanation for the difference might be that urban clients are more likely to transfer providers because they have more access to information and other options. Furthermore, metropolitan customers may be more receptive to trying out new goods and services since they often have free time to spend cash.

The greater percentage of customer churn in metropolitan regions suggests that keeping these customers is more difficult. It is important that the company determines and addresses the precise causes of this high incidence of customer churn. Retention tactics that are specifically designed to meet the demands and tastes of urban customers may be crucial. On the other hand, even while the non-churn rate for rural customers is somewhat lower, there is still a great deal of space for improvement when it comes to customer retention. Positive outcomes may arise from developing ways to improve customer satisfaction and meeting the unique needs of rural clients.

4.3.6 Technology and Network

In this phase, it will be examined whether the technology feature has an impact on churn or not.

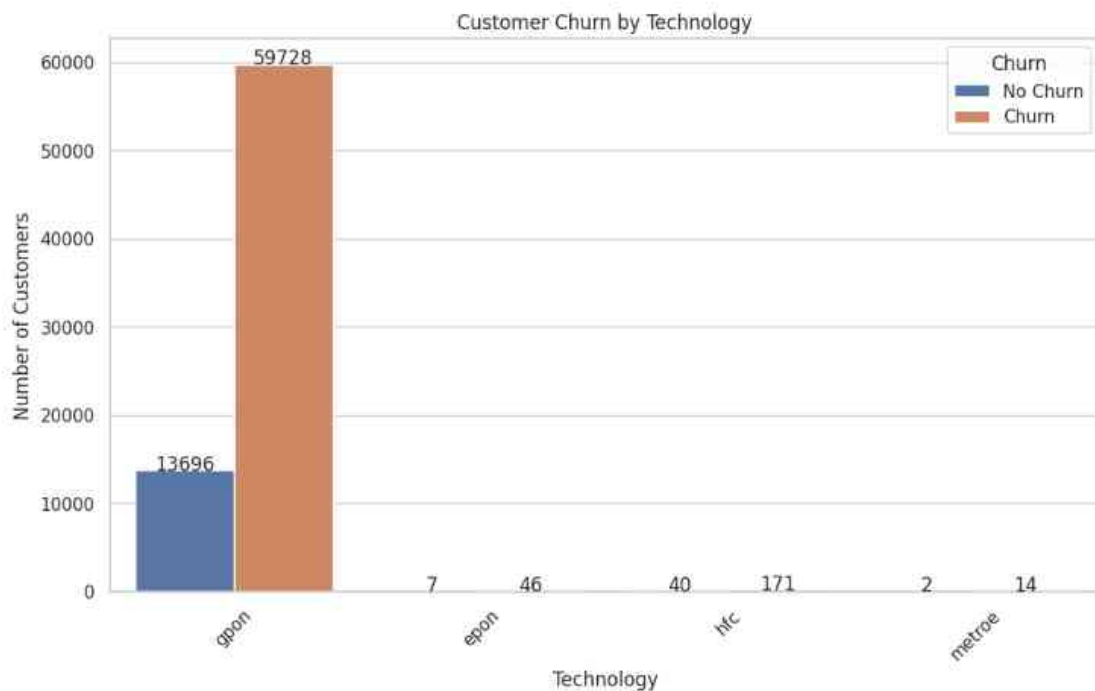


Figure 4.29 Customer Churn by Technology

The statistical test:

Chi-Square Statistic: 1.4472790398548632

p-value: 0.6944906331563265

The plotting of customer churn across several technologies at PT XYZ reveals that the most customers, both churned and not churned, are associated with GPON technology, which is the newest in optical fiber. This is consistent with PT XYZ's resolve to implement advanced technological infrastructure. Since this advanced technology may greatly increase customer satisfaction and loyalty, it is essential to educate clients about how it improves their online experience.

Despite the advancements in technology, statistical research reveals no discernible correlation between the types of technology, such as GPON, and customer churn. This finding implies that while technology plays a significant role in PT XYZ's service offering, it may not be the only element affecting the retention of customers.

To lower churn, PT XYZ is encouraged to concentrate on other important factors. These include providing value-added services, enhancing the client experience and service quality, and broadening the marketing plan beyond technology. Crucial components include offering competitive pricing, providing excellent customer service, and developing customized marketing strategies that relate to the requirements and interests of the customers that the company is targeting.

In conclusion, PT XYZ finds that a comprehensive strategy that smoothly combines superior customer service, ongoing innovation, and satisfaction with advanced technology is a more successful way to keep customers. This all-encompassing strategy may assist in improving overall customer loyalty and addressing the more general components of consumer expectations.

4.3.7 Customer Account Information

There are three features in this category: *cust_seg*, *prepaidtype*, and *tik_custid*. Since *tik_custid* is associated with the customer ID, further analysis will not be conducted. Having already created a client segment based on *prepaidtype*, the analysis will now focus solely on examining the *cust_seg* attributes and their correlation to churn.

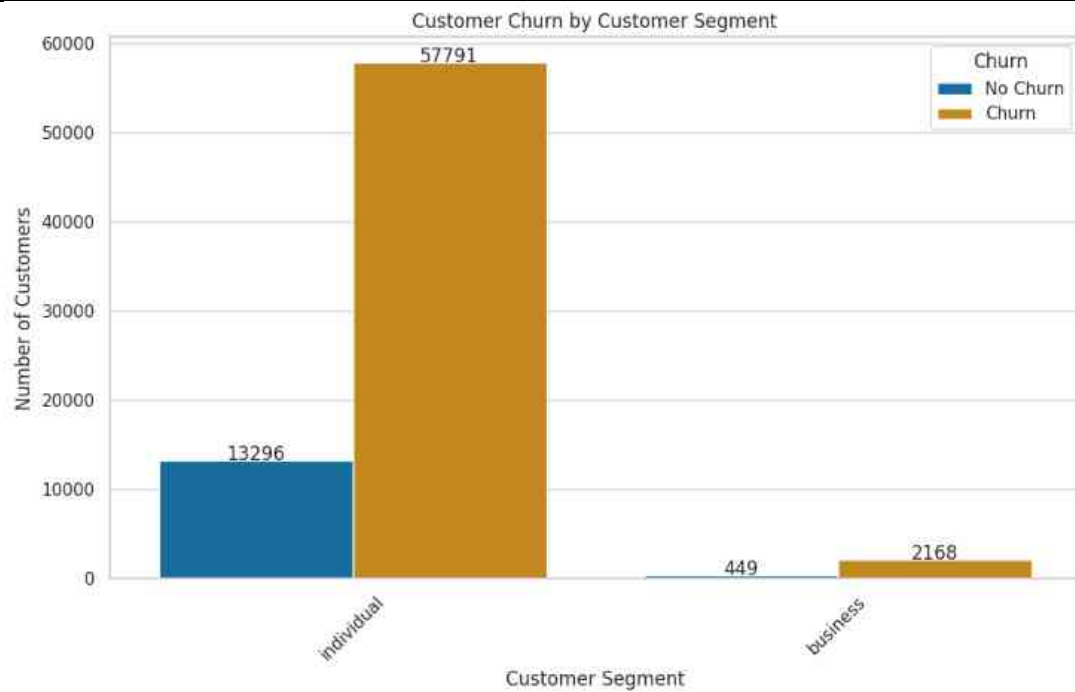


Figure 4.30 Customer Churn by Customer Segment

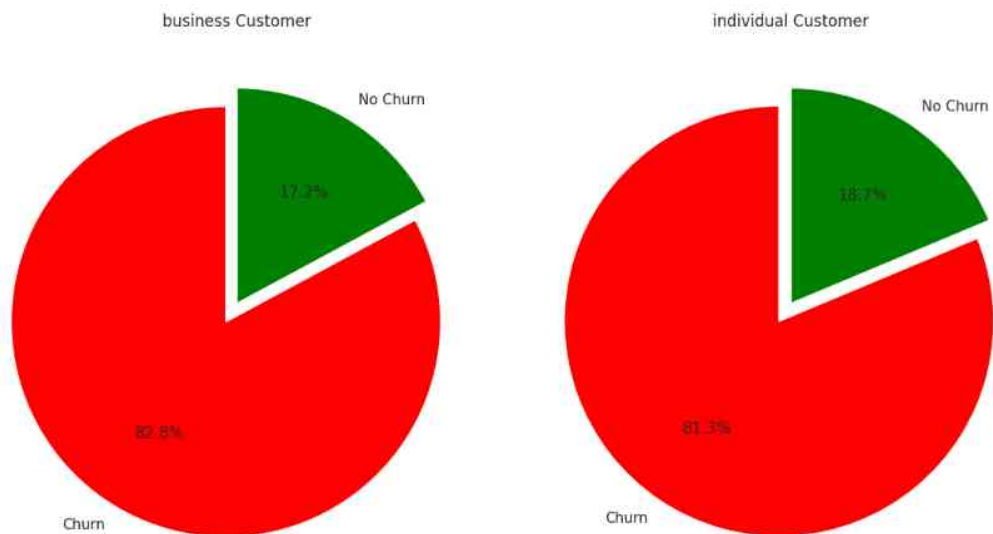


Figure 4.31 Customer Churn Rate Percentage by Customer Segment

The statistical test:

Chi-Square Statistic: 3.8793053398553883

p-value: 0.04888477964774258

The graph and pie chart analysis shows how PT XYZ's churn is allocated across its various customer categories. This distribution shows that various consumer categories have different levels of engagement or satisfaction with the services offered. To better serve the specific needs of each customer and increase retention, tactics including increasing customized communication, emphasizing customer service quality, and putting loyalty programs or incentives into place may be essential.

On the other hand, creating specialized business solutions and packages that satisfy their needs, developing stronger relationships with customers via proactive account management, and providing scalable services are probably going to be more successful approaches for corporate customers. These strategies may support long-term collaborations while meeting the changing demands from business customers.

Further assisting PT XYZ in improving its customer engagement and retention tactics is the ability to get deeper insights into the various wants and preferences of its customers via the administration of customer satisfaction surveys.

Crucially, the statistical test shows that there is a statistically significant correlation between turnover and client categories. This implies that a customer's categorization as an individual or a business customer influences, at least in part, the chance of the customer churning. PT XYZ should therefore take these segment-specific variations into account when managing client interactions and creating focused churn reduction initiatives.

4.3.8 Target Variable

The final analysis to be conducted will focus on the target variable itself, *churn*.

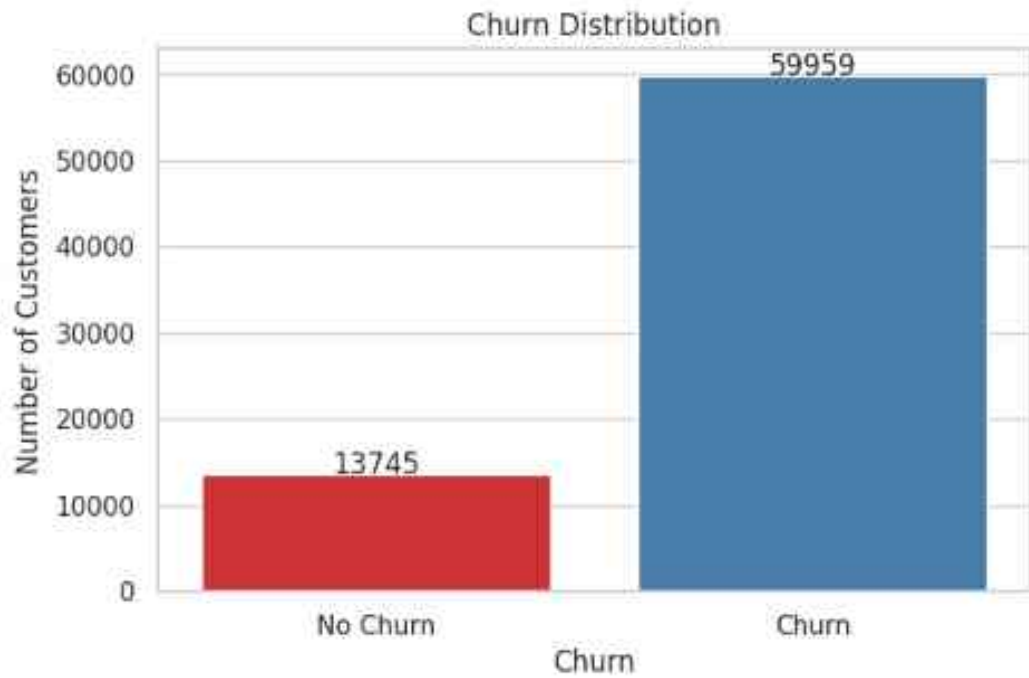


Figure 4.32 Churn Distribution

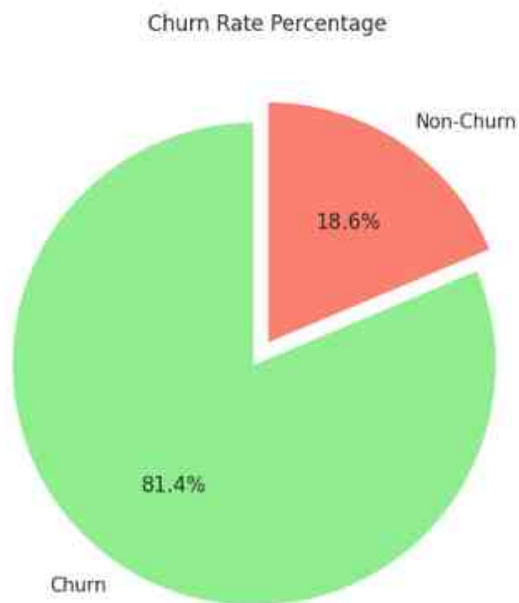


Figure 4.33 Churn Rate Percentage

Figures 4.33 and 4.34 indicates an imbalance and a significant gap between the number of consumers who have churned and those who have not churned in the sample, which consists of 73,706 customers. The churn rate is much higher, with 59,960 out of the total 73,706 subscribers discontinuing their service. Approximately 81.4% of the overall client base ended their services with the Internet Service Provider, indicating a significant majority. On the contrary, the number of consumers that did not churn is far less, including just 13,746 out of the overall 73,706 customers. It represents around 18.6% of the whole customer base, suggesting that there is a lower percentage of consumers who have retained their services with the Internet service provider without cancelling. This observation indicates a concerning problem with client retention for the Internet service provider. The substantially higher turnover rate suggests possible concerns such as dissatisfaction with services, price, customer support, or the availability of better options elsewhere. It is essential for the ISP to identify and resolve the underlying factors contributing to this significant customer turnover rate in order to enhance customer happiness, retention tactics, and overall company performance.

4.4 Model Design

Before beginning the creation of an MLP churn prediction model, it is essential to select features based on the Exploratory Data Analysis (EDA) process and ensure that they are the most suitable for use in the model.

4.4.1 Feature Selection

Table 4.13 displays the features that have been selected for modeling, based on the examination conducted during the Exploratory Data Analysis (EDA) process.

Table 4.13 Feature Selection for Modeling

Features Group	Features
Categorical Data	packagename, call_status, category, sa_province, territory, cust_seg, area_type
Numerical Data	packageprice_idr, subscription_tenure, total_tickets
Target	churn

It has been verified that there are no missing values in the selected features from the initial phase of the feature selection process, indicating that the dataset is complete for these variables. The data types—numerical characteristics of types int64 and float64 and categorical features of type object—are suitable for the modeling assignment. Following this, the next step involves moving on to feature correlation and conducting statistical analysis of the relationship between the target and the selected features.

4.4.2 Feature Correlation

In this phase, the correlation between numerical features and the target will be checked using a correlation matrix.

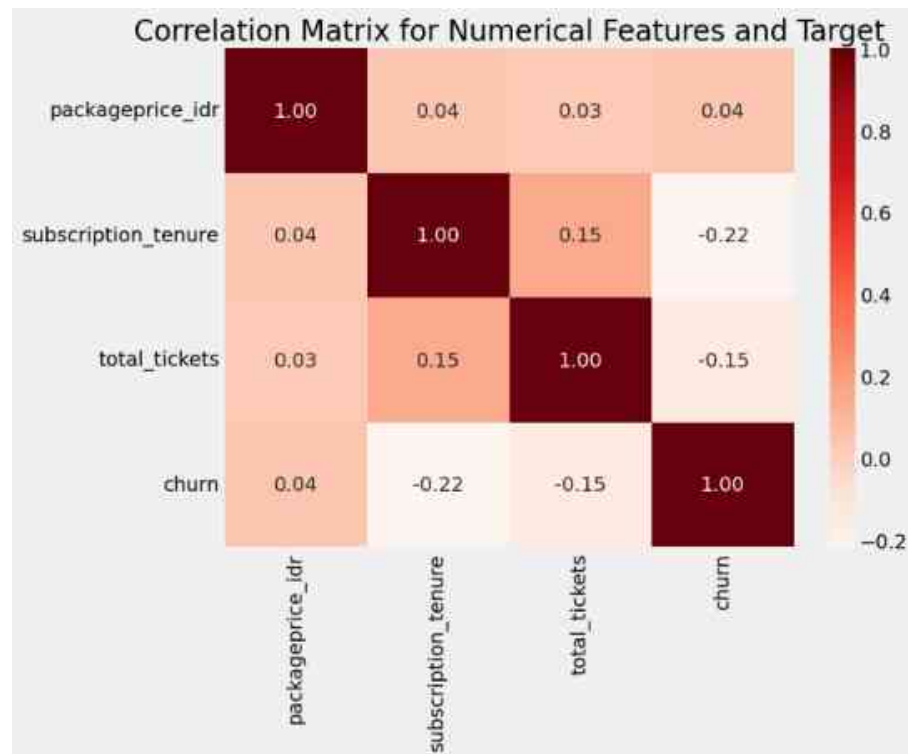


Figure 4.34 Correlation Matrix for Numerical Features and Target

Based on the correlation matrix, an analysis can be conducted on the numerical features in relation to the target feature (churn):

- *subscription_tenure* has a negative correlation with churn (-0.22), suggesting that longer subscription periods may be associated with lower rates of customer churn.

- *total_tickets* also have a negative correlation with churn (-0.15), reflecting that a smaller probability of churn may be associated with a larger quantity of tickets.
- *packageprice_idr* shows a very weak correlation with churn (0.04), it suggests that, on its own and based on the analysis on EDA, package pricing although may not be a reliable predictor of turnover but still can provide comprehensive view of customer behavior.

Based on the feature correlation test, it shows that there are no strong correlations between the variables presented in the matrix. An ANOVA (Analysis of Variance) test will be conducted to ensure that the numerical characteristics may be used for the models.

Table 4.14 Numeric ANOVA Test

index	Feature	ANOVA Score	P-Value
1	subscription_tenure	2.970,95	0.0
2	total_tickets	1.289,55	2,22E-264
3	packageprice_idr	94,51	2,54E-06

Based on out ANOVA test, it shown that:

- *subscription_tenure*: a high ANOVA score and low P-Value (less than 0,05) indicates that *subscription_tenure* has a strong relation with churn. This might suggest that consumers are more likely to churn after a certain amount of time, or that newer customers are more likely to churn.
- *total_tickets*: the result was same with *subscription_tenure*, the ANOVA and P-Value indicates that *total_tickets* have a significant effect on churn. A customer with a high ticket might indicate dissatisfaction or issues leading to churn, or it might suggest more engagement.
- *packageprice_idr*: although a low ANOVA score shows that even it has a less obvious relation with churn, a very low P-Values shows that *package_price* statistically has a significant effect on churn.

Based on the ANOVA test, it is concluded that *subscription_tenure*, *total_tickets*, and *packageprice_idr* will be the features in churn prediction model.

4.4.3 Statistical Testing

After completing the correlation analysis for numeric features, the next step will involve performing a statistical test for categorical features. In this phase, a chi-square test will be conducted using label encoding and a random train-test split in the process.

Table 4.15 Chi-Square Categorical Result

Index	Feature	Chi-Square	p-value
0	packagename	2.878,78	0.000
1	call_status	22.680,88	0.000
2	category	42.805,14	0.000
3	sa_province	304,69	8.92e-53
4	territory	7,07	0.0078
5	cust_seg	0,99	0.3196
6	area_type	33,09	8.79e-09

Based on table 4.15, it shown that features like *packagename*, *call_status*, and *category* have very high Chi-Square values and low p-values which indicating a very strong statistical association with the target feature churn. *sa_province*, *territory*, and *area_type* features also show significant associations, however not as much as the features stated above. Meanwhile, *cust_seg* has a higher p-value which suggest a weaker association with the target feature.

Based on the chi-square test, *packagename*, *call_status*, *category*, *sa_province*, *territory*, and *area_type* will be used as the features in churn prediction models.

4.4.4 Baseline Data Preprocessing

Next, baseline data preprocessing will be conducted, implementing a standard procedure for preparing the data for modeling. Initially, the data will be divided into distinct subsets—training, validation, and testing sets—to facilitate model development, parameter tuning, and evaluation. Two fundamental transformations were employed:

- Normalization for numerical columns (*subscription_tenure*, *total_tickets*, and *packageprice_idr*) to ensure the features are on the same scale.

- Encoding for categorical columns (*packagename*, *call_status*, *category*, *sa_province*, *territory*, and *area_type*) to convert the categorical features into a format that is suitable for machine learning models.

To address the imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) will be used. By generating synthetic samples for the minority class (which might be underrepresented), SMOTE created a more balanced representation of classes within the training data. This ensured that the model could learn from a more equitable distribution of classes, potentially enhancing its ability to generalize and make accurate predictions, particularly in scenarios where class imbalance might adversely affect model performance.

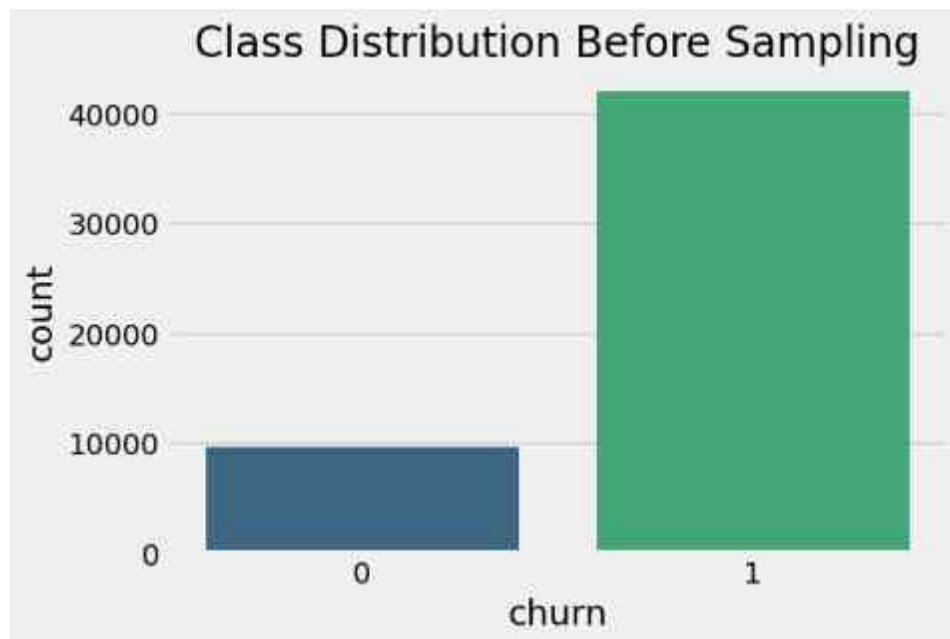


Figure 4.35 Class Distribution Before Sampling (SMOTE)

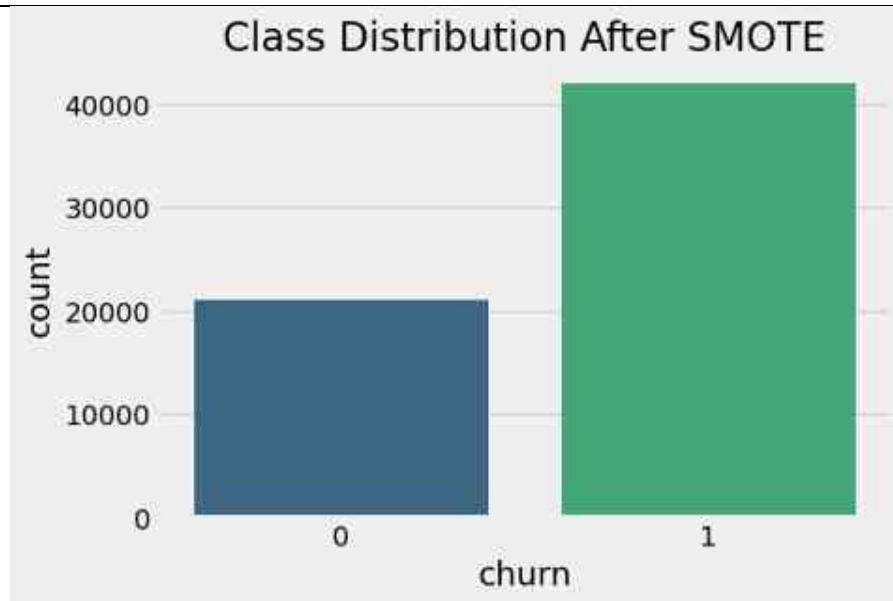


Figure 4.36 Class Distribution After SMOTE

Figures 4.35 and 4.36 show the results before and after the process of handling imbalance data. The train data has been adjusted to 50% of the majority class. By doing this, the model's performance in predicting non-churn customers will improve.

4.4.5 Model Development

In this phase, an initial bias will be set in the model development for a Multi-Layer Perceptron (MLP). This procedure may be very important, particularly in binary classification issues when unbalanced datasets are present. Because the majority class is more frequently in the training data, the model may initially tend to predict it more frequently in situations when the classes are substantially unbalanced (one class considerably outnumbers the other). This causes a bias in favor of the dominant class, which makes it difficult to identify the minority class accurately.

Establishing an initial bias gives the model a starting position that pushes predictions in the direction of a more balanced viewpoint, which helps overcome this tendency. The model begins with a better knowledge of the underlying class distribution by initializing the bias to a value that approximates the class distribution (such as the logarithm of the ratio between the positive and negative classes). As a result, during the early training phases, it is less likely to be overly biased towards the majority class.

Initial bias result: 1.47725254

The initial bias calculation reveals a significant imbalance in the dataset, with a higher occurrence of churned customers compared to non-churned ones. By setting the initial bias in the model, the model's starting predictions are adjusted to better reflect the actual class distribution. This approach is particularly useful in mitigating the model's initial inclination to underpredict the positive (churn) class, especially during the early stages of training. Implementing this initial bias is a strategic step towards addressing class imbalance, ensuring that the model does not overlook the important, albeit less frequent, instances of customer churn.

After processing with the initial bias, the model can be tested and compared when run to predict churn with and without the initial bias. The result of this process is shown on table 4.16.

Table 4.16 Model Test with and Without Initial Bias

	Without Initial Bias	With Initial Bias
Model Prediction	0,76	0,85
	0,73	0,85
	0,74	0,84
	0,72	0,85
	0,73	0,86
	0,74	0,87
	0,73	0,86
	0,75	0,86
	0,73	0,87
	0,75	0,83
Model Evaluation (Loss)	0,51	0,47

Predicted probabilities, or the model's estimated probability of a sample belonging to the positive class, are produced by the model for the positive class in a subset of training data. The model was trained on a random data set, and the probabilities are based on the model's present state. Due to the randomness of the training process, there may be insignificant fluctuations in predictions; nevertheless, they are a normal component of the learning process of the model and have nothing to do with the bias initialization of

the output layer. In unbalanced data, the initial bias is mainly utilized to match the model's early predictions with the real class distribution.

The average loss per sample across the whole training dataset is represented by the loss value. This measure, which for binary classification tasks is usually a binary cross-entropy loss, indicates how well the model is doing, with lower values indicating greater performance. But the relevance of this loss value also relies on how complicated the dataset is and what unique difficulties the prediction job presents.

In conclusion, the loss value gives a more comprehensive picture of the model's overall performance, while the prediction array offers insights on the model's present grasp of the data. To completely evaluate and enhance the model's churn prediction efficacy, further model assessment is required. This evaluation should include validation and testing on other datasets as well as consideration of new metrics.

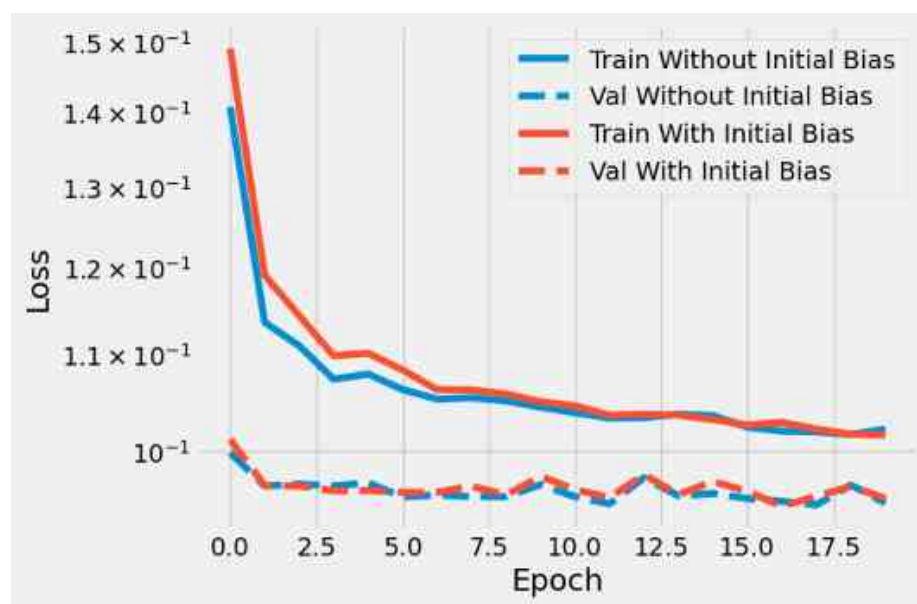


Figure 4.38 Train and Validation Model Test with and Without Initial Bias

Figure 4.38 shows the training and validation loss across several epochs for two scenarios, one in which the model has an initial bias and the other without. In the beginning, the training and validation datasets have a greater loss for the model with initial bias. Both models demonstrate a declining trend in loss with an increase in epochs, which suggests that learning is progressing. The fact that the losses in both

cases eventually approach the same levels indicates that the initial bias's influence decreases with enough training. The end-of-training validation loss reflects this convergence, suggesting that the initial bias may not have a substantial impact on the model's capacity to generalize. This condition will be demonstrated when a test is conducted on the actual model using the initial weight.

Next, the determination of how many layers and neurons will be used for the model is for the MLP model:

- 1 input layer with 9 neurons. 9 selected features will be the input layer.
- 2 hidden layers with 64 neurons on the first layer and 32 neurons on the second layer.
- 1 output layer with 1 neuron, which is a churn feature.

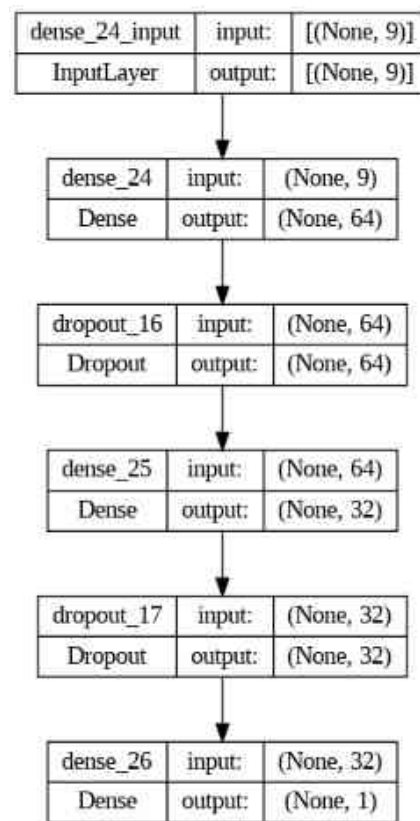


Figure 4.39 MLP Model Architecture

Applying class weights is an essential next step after creating the models, especially when working with unbalanced datasets. This method involves providing the minority

class with a higher weight and the majority class a lower weight. The model is informed that mistakes in the minority class are more significant than those on the majority class by using these class weights to affect the loss function during training.

The class weights result:

- Weight for class 0: 2.69
- Weight for class 1: 0.61

Class 0, which is the minority class, has a weight of 2.69. This means that the model's learning process will be impacted by each occurrence of class 0, which is equal to almost 2.69 instances of an equally distributed class. On the other hand, class 1, which is the majority class, has a lower weight of 0.61, which represents its greater frequency in the dataset.

By using these class weights in the training phase, the loss function is adjusted, increasing the sensitivity of the model to mistakes on the minority class. The objective of this technique is to get a more accurate representation and maybe better performance in both classes by balancing the effect of each class on the model. After completing all the necessary preparations for modeling, the next step is to train and test the model.

4.4.6 Model Train and Test

When creating accurate and effective machine learning models, particularly for classification tasks, the treatment of the dataset can significantly impact the model's performance. This is particularly important in cases when there is a possibility of imbalance in the dataset, which favors one class over another. To investigate the effects of various data handling strategies on model performance, the model will be trained and evaluated in three different scenarios:

- Using the Original Data: The original dataset will be used to train and test the mode. This method will act as a starting point, giving information on the model's performance in the absence of any adjustments to correct for any class imbalances.

- **Applying Class Weights:** In this scenario, the class imbalance is addressed by giving each class an independent weight during the training phase. This method aims to make the model more sensitive to the minority class by increasing the cost of misclassification for that class. It is an effective way to improve model performance on datasets that are not evenly distributed.
- **Using the Synthetic Minority Over-sampling Technique (SMOTE):** This method may help the model generalize to a wider range since it balances the class distribution and adds variability to the dataset. The goal in training the model using SMOTE-augmented data is to find out how well over-sampling works to improve model performance when there is a class imbalance.

Table 4.17 Model Performance Results

Model Performance Metrics	Performance Result		
	Using Original Data	Applying Class Weight	Using SMOTE
Loss	0,11	0,15	0,11
Precision	0,99	1,00	1,00
Recall	0,96	0,94	0,95
AUC	0,98	0,98	0,99
F1 Score	0,97	0,97	0,97
Confusion Matrix	tp : 8636.0 fp : 98.0 tn : 1946.0 fn : 376.0	tp : 8477.0 fp : 5.0 tn : 2039.0 fn : 535.0	tp : 8542.0 fp : 12.0 tn : 2032.0 fn : 470.0

Table 4.17 shows the comparison results of the model performance by applying three different conditions. It shows that:

- **Using Original Data.**
High accuracy and recall are shown by the model trained on the original data, resulting in a high F1 score of around 0.973. Additionally, a very high AUC (Area Under the Curve) indicates great model performance. But in contrast to the true positives (tp), the true negatives (tn), which may indicate a class imbalance.
- **Applying Class Weight.**

Although there is a little increase in loss when class weights are used to address imbalance, accuracy remains excellent. Recall shows a noticeable decline, meaning that while the model is still accurate, it is no longer as good at identifying all positive data. The somewhat lower F1 score of 0.969 compared to the original data model might be the result of a trade-off between recall and accuracy. On the other hand, there have become a lot of true negatives, indicating that the capacity to recognize negative situations has improved.

- Using SMOTE.

The loss with SMOTE-augmented data is similar to that of the original data case. With precision and recall somewhat better than in the class weight scenario, the greatest F1 score—roughly 0.973—is obtained. The fact that the AUC is the greatest of the three indicates that the genuine positive rate and false positive rate are most favorably balanced. While there have been more false negatives than in the original data, the model has produced better genuine negative predictions by balancing the trade-off between various sorts of mistakes.

Based on the analysis on the three conditions of testing, SMOTE appears to offer a good balance between sensitivity and specificity, as indicated by the highest F1 score and AUC, suggesting that the model is more powerful to class imbalance.

4.5 Model Evaluation

After the training and test procedure is carried out, it is concluded that the best model for predicting churn is the one with SMOTE.

Table 4.18 SMOTE Performance Metrics

SMOTE Performance Metrics	
Loss	0,11
Precision	1,00
Recall	0,95
AUC	0,99
F1 Score	0,97

Based on SMOTE performance metrics result, it shown that:

- The loss showing a good value which measure of how well the model is performing from a probabilistic perspective.
- It has a high precision value which indicates when the model predicts an instance is positive, it is correct almost all the time.
- The recall value is also high, suggesting the model can identify a high percentage (95%) of actual positive (churned) instances.
- The area under the ROC (AUC) curve is close to 1, this indicates that the model does a great job at differentiating between the classes.
- The F1 Score, which is the most important metric in an imbalanced classification setting, has a high value. This condition indicates a balance between precision and recall.
- Based on recall value, the model with SMOTE can correctly identify 95% of churn cases. It indicates that the model is quite effective at identifying customers who are likely to churn.

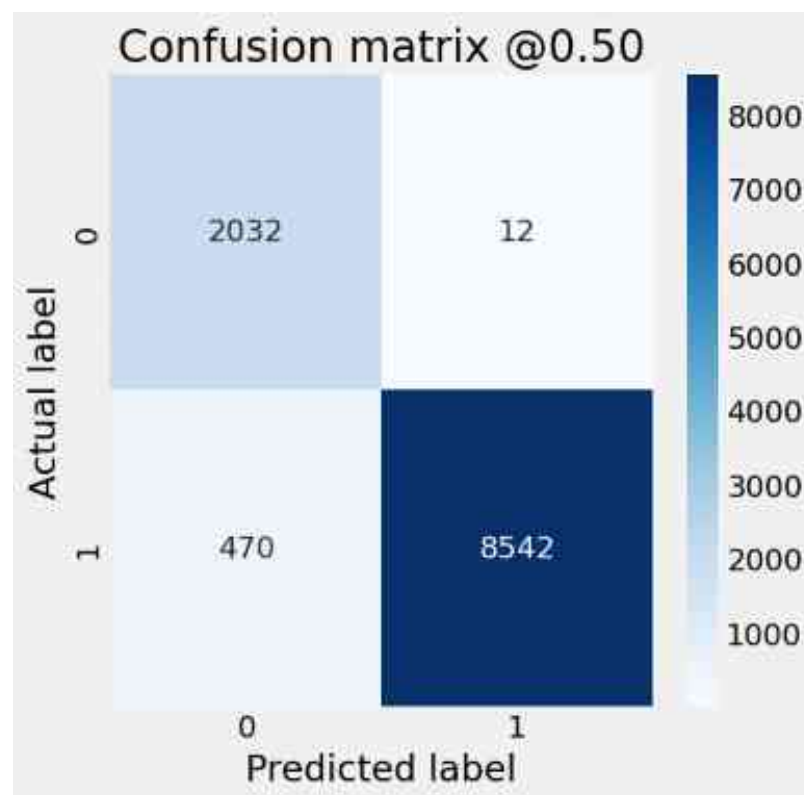


Figure 4.40 SMOTE Confusion Matrix Result

Based on SMOTE confusion matrix result with 9012 test data, it shown that:

- True Positives (TP) - 8542: the model correctly identified 8542 instances as positive (churned customers) which means the model has good sensitivity.
- True Negatives (TN) - 2032: the model correctly identified 2032 instances as negative (non-churned customers) which means the model has a good specificity.
- False Positives (FP) - 12: only 12 non-churned customers were incorrectly identified as churned which means the model is contributing to high precision.
- False Negatives (FN) - 470: the model missed 470 churned customers, labeling them as non-churned. While higher than FP, this number is relatively small compared to the number of true positives.

Now, analyze what business implications can be derived from the predictions made by this model.

- True Positives (TP).
This group comprises customers correctly identified as being at risk of churning. Timely and effective interventions can potentially save these customers, retaining revenue and preserving customer relationships. The model's success in identifying a high number of true positives is beneficial for targeting retention efforts.
- True Negatives (TN).
These are the customers that the model correctly predicted would not churn. No action is required for these individuals, which helps in allocating the business's retention resources more efficiently. The high number of true negatives suggests the model is effective in avoiding unnecessary retention actions.
- False Positives (FP).
These represent the customers who were incorrectly predicted to churn. Any intervention, such as offering discounts or special incentives to these individuals, could result in unnecessary expenditure for the company. However, since the model has very few false positives, the cost associated with this error is minimal.
- False Negatives (FN).

These are the customers who were predicted to stay but churned. This category poses a more significant concern as it directly equates to lost customers and, consequently, lost revenue. It's critical to minimize this group as much as possible since identifying these customers correctly could lead to successful retention strategies.

By focusing on these interpretations, businesses can allocate their resources more effectively, targeting customers who are most at risk of churning according to the model's predictions, while avoiding unnecessary costs on those who are likely to stay.

4.6 Integration of MLP and EDA

Several crucial conclusions can be drawn based on the comprehensive EDA analysis and the outcomes of the predictive modeling using SMOTE. The requirement for focused retention tactics found in the EDA is consistently met by the model's accuracy in forecasting both non-churn (True Negatives) and churn (True Positives). Due to this synergy, the actions can be more accurately targeted towards clients who are most likely to churn.

The implications for True Positives (TP) are clear. The EDA emphasized how crucial it is to customize campaigns according to specifics such as age, subscription information, and regional differences. When developing individualized retention tactics for clients who have been recognized as at-risk, these insights are quite helpful.

The True Negatives (TN) outcomes align with the EDA's findings, which suggested that not all demographics or customer segments necessitate intensive churn prevention strategies. Gaining this insight will help to allocate the retention resources more wisely and concentrate the efforts where they are most needed.

The implications of False Positives (FP) are also significant. The model's low incidence of false positives highlights the need to effectively allocating marketing resources and finances, a tactic that the EDA determined to be essential. By doing this, needless spending on clients who are unlikely to defect is reduced.

False Negatives (FN) present a critical area for ongoing attention. This feature emphasizes the need of continuously improving the churn prediction model in order to better identify minor signs of churn risk, by adding more nuanced data points as recommended by the EDA.

To sum up, the integration of the EDA with the MLP prediction model provides solid groundwork for building a sophisticated client retention approach. Personalized engagement and retention tactics may be informed by demographic trends, activity patterns, and consumer insights obtained from the EDA. In the meanwhile, the model's predictive accuracy makes it possible to target these techniques precisely at the people who are most vulnerable, guaranteeing efficient resource allocation and improved churn control.

Based on the integration of EDA and MLP, some business example scenarios can be simulated.

1. Cost of False Positives (FP).

This would include the cost of discounts/offers given to customers who wouldn't have churned. Calculate the average cost per intervention and multiply it by the number of false positives. Considering the cost of offering a retention package or discounts to customers who are predicted to churn but wouldn't have otherwise. Assume an average cost of 300,000 IDR per intervention.

- $\text{Cost of FP} = \text{Cost per Intervention} * \text{Number of False Positives}$
- $\text{Cost of FP} = 300,000 \text{ IDR} * 12 = 3,600,000 \text{ IDR}$

2. Cost of False Negatives (FN).

This is the revenue lost due to customers churning that the model failed to identify. Calculate the lifetime value of a customer and multiply it by the number of false negatives. This could be the average monthly revenue per customer multiplied by the number of months a customer typically stays.

Suppose the average monthly revenue per customer is 400,000 IDR, and the average customer stays for 24 months.

- Revenue Lost per Churned Customer = Monthly revenue per customer *
Number of months
- Revenue Lost per Churned Customer = 400,000 IDR * 24 = 9,600,000 IDR

Now, calculate the total revenue lost due to False Negatives:

- Cost of FN = Revenue Lost per Churned Customer * Number of False
Negatives
- Cost of FN = 9,600,000 IDR * 470 = 4,512,000,000 IDR

3. Revenue Saved from True Positives (TP).

Calculate the potential revenue retained by applying interventions to these customers. Multiply the average retained revenue per customer by the number of true positives. Assume successful interventions can retain a customer for an additional 12 months, resulting in an additional revenue.

- Estimated additional revenue per retained customer = Monthly revenue per
customer * Number of additional months retained
- Estimated additional revenue per retained customer = 400,000 IDR * 12 =
4,800,000 IDR

Now, calculate the total revenue retained from True Positives:

- Revenue from TP = Estimated additional revenue per retained customer *
Number of True Positives
- Revenue from TP = 4,800,000 IDR * 8542 = 41,001,600,000 IDR

4. Savings from True Negatives (TN).

These are customers correctly identified as not churning. No intervention is needed, so there's a cost-saving here.

Conclusion:

- Total Cost = Cost of FP + Cost of FN
- Total Revenue = Revenue from TP

Total Cost = 3,600,000 IDR + 9,600,000 IDR = 13,200,000 IDR

Total Revenue = 41,001,600,000 IDR

- The revenue generated from successful interventions (TP) significantly outweighs the costs incurred due to false positives and false negatives.
- Implementing effective interventions based on churn prediction results in substantial revenue retention, which surpasses the costs of offering interventions to some customers who might not have churned.

In summary, the business calculation suggests that while there are costs associated with misidentifying churn (FP and FN), the revenue generated from accurate predictions and successful interventions (TP) far exceeds these costs. Therefore, the churn prediction model, despite some margin of error, proves to be beneficial in retaining revenue by identifying and successfully addressing potential churn cases.

4.7 Findings

On this subsection contains about a summary of all the processes performed on chapter 4. The explanations given in this subsection expand further basic facts to offer information on the underlying reasons in customer churn and provide possible solutions for PT XYZ. The purpose of this subchapter is to explain these results by providing a thorough and clear examination of the major trends and revelations that have resulted from the investigation.

1. On subsection 4.1, the dataset collected from PT XYZ contains mix characteristics of categorical and numerical variables that offers an insight into customer behaviors. The dataset already provided demographic information, subscriptions details, billing / payment information, customers interaction, geographic information, technology, customer account information, and the target in the research, which is churn.
2. On subsection 4.2, data preparation and preprocessing were conducted. These steps were crucial in ensuring the reliability and effectiveness of subsequent analysis and modeling. Handling with data duplication, missing values, outliers (especially in customer age), selecting variables, and transform data for numerical and

- categorical data into a format suitable for machine learning, help to reduce computational time, improve prediction performance, and enhance data quality.
3. On customer demographic information, Figure 4.10 shows a marginal difference in churn rates between male and female. Although statistically customer gender was not significant related with churn, gender-specific preferences and behaviors might still be relevant for targeted marketing and customer service strategies. On the other hand, Figure 4.11 shows a group of customer age that's become a significant factor in churn and retention. Although customer age was not become the features in modeling because of the outliers, tailored marketing and retention strategies that are specifically designed were crucial to address the unique needs and expectations of different age groups, especially customer age group 25-34.
 4. Subscription details show finding in relation between package price and subscription tenure. On package price, it shows that customers who starts with higher price are more likely to churn (25th percentile (Q1)). This might point to a situation in which consumers who choose more costly packages join the service with higher expectations or a greater sensitivity to the perceived value they get. As for subscription tenure, Figure 4.15 shows that customers in the early stages of their subscription with PT XYZ are more likely to churn. Welcoming experiences, early engagement programs, and attractive introductory offers in this initial phase become important to raise customer loyalty.
 5. Billing and payment information subsection highlights the challenges and limitations of working with incomplete datasets, especially in understanding complex aspects such as billing and payment information. Given only information about branch name and lacking detailed elements such as payment method, detailed billing information, and payment frequency, makes the analysis cannot provide a deeper insight into the dynamics of billing and payment.
 6. Based on customer interaction with the company, figure 4.17 shows that customers who churn tend to have fewer tickets compared to those who do not churn. This condition implies that efficient communication and proactive customers service are key to improving customer retention. Figure 4.18 until 4.21, shows the relation of customers churn which most comes from the inability in contacting customers. With top five reasons why contacting customers becomes failed (call failed),

- indicates that there are a need for some improvement or another media in contact with customers. Some possibilities reason that can cause these conditions are calling at odd hours, unidentified caller IDs, and customer hesitancy to answer unknown numbers.
7. For geographic information, the findings indicate that geographic location significantly influences customer churn, with variations observed across provinces, territories, and urban/rural area. These conditions underscore the necessity of PT XYZ to develop customized marketing campaigns with specific regions based on their churn rate condition and customer's needs.
 8. As PT XYZ is already implementing the newest technology in fiber optics, Figure 4.29 shows that PT XYZ want to give their customers with the newest technology.
 9. In the customer account information subsection, focuses on the attribute related with customer segmentation (individual or business) which emphasizes the correlation between customer segmentation and churn. Figure 4.30 shows the significance of using segment-specific methods in customer relationship management and creating focused plans to lower churn rates. Different strategies and approaches need to be developed by PT XYZ to handle individual and business customers. Conducting a customer survey can give an insight into what kind of service or product those segments need.
 10. Based on the result from EDA, 10 selected features were used for the modeling. The features were chosen related with their correlation with churn based on statistical testing.
 11. Subsection feature correlation conduct to make sure if the numerical category has a strong correlation with churn by using the train and test dataset. The correlation matrix shows a weak relation between the numerical category with churn. On the other hand, second test using ANOVA, shows that based on the p-value (the alpha number for p-value is 0,05) indicate there are a positive correlation between numerical category with churn.
 12. Statistical testing is conducted to find a correlation between categorical variables with churn. In this subsection, the variables down to 9 variables where customer segmentation do not show any strong correlation with churn as the target based on p-value result.

13. The baseline data preprocessing involved standard procedures to prepare the data for modeling, ensuring that the selected features were properly encoded and scaled for use in machine learning algorithms. Two fundamental transformations were employed in these subsections: normalization for numerical columns and encoding for categorical column. SMOTE technique was conducted to handle an imbalance data from the dataset to create a more balanced representation of classes.
14. In model development, initial bias was set to gives the model a starting position that pushes predictions in the direction of a more balanced viewpoint. 1 input layer with 9 neurons (representing 9 selected features), 2 hidden layers each with 64 neurons and 32 neurons, and 1 output layer with 1 neuron (churn variable), were set up as the basis modeling in MLP. Class weight also applied in this subsection after deciding the model architecture. By applying class weights, the model is informed that mistakes in the minority class are more significant than those in the majority class.
15. Table 4.17 shows the result in model train and test subsection. The table shows a model performance comparison result from three scenarios. The first scenario, which used original data, shows an accuracy of 96% (based on recall value) in predicting churn. The second scenario, which used class weight in, shows an accuracy of 94%. The last scenario using SMOTE, shows an accuracy of 95%. Based on the scenarios in handling the imbalance data, the model with SMOTE technique appears to offer a good balance in sensitivity and specificity indicated by the highest F1 and AUC score.
16. In the model evaluation using SMOTE, performance metrics of SMOTE model were highly effective in predicting churn, striking a balance between identifying churned customers accurately and minimizing false positives. Figure 4.40, which shows SMOTE confusion matrix result, strengthen the model evaluation using SMOTE. The confusion matrix indicates a well-balanced model with effective capabilities in correctly classifying both churned and non-churned customers. The low number of false positives and high true positive rate highlight the model's accuracy and reliability in predicting customer churn.

17. In the last subsection, the integration of EDA and MLP represents a comprehensive approach to understanding and addressing customer churn, leveraging both data exploration and predictive modeling to inform and refine client retention strategies.

4.8 Deployment

Having developed a churn prediction model with an impressive accuracy of 95%, the point at which this model can be implemented in a real setting is now being approached. However, the deployment procedure is more complicated than it first seems. There are several phases involved, each with unique needs and obstacles. Every stage is essential to ensure that the model functions as best it can in a real-world scenario, from model validation to application integration, from UI creation to ongoing monitoring and updating.

To integrate the model into the system in PT XYZ, several important point should be taken. Ensuring there are no missing and duplicate data on customer information (customer age, subscription tenure, total tickets, package price, and package name) and geographic information (province, territory). Making a new variable that meet the requirement of the model also become a crucial part in deployment. Selecting how to integrate with the existing system, modeling update, monitoring model performance and evaluation to meet the expectations, and the approval from the company management becomes the important point.

When all the requirement for deployment has been meet, the model will become a powerful tool in detecting the possibility of customer in churning. The models can be run for customers who are entering the grace period of their subscription. Being able in predicting customer who has the possibility in churn, PT XYZ can make a marketing strategy based on the result from the models. A retention campaign that related with customer age group or territory can be offered to those customers.