Churn Prediction for Telecom Customers

Moh. Ridwan

1. Summary

Business Background

PT X is a company engaged in telecommunications. In the previous month there were consumers who left the company's services.

Problems Statements

How can PT X maintain the number of consumers who use its services ?

Objective

Analyze the data and predict the churn of us ers (to identify people who will and will not re new their contract)

Proposed Solutions

PT X can make a model that can predict the classification of loyal consumers or not based on the data. After that, PT X can focus on the prediction results of disloyal consumers and create a model to cluster these consumers.

Result:

The classification of disloyal consumers can apply the logistic regression model with recall = 81.41% and roc-auc score = 76,23%. Some of the recommendations generated:

- There needs to be more attention to new users, for example in the form of discounts or raffle points.
- Promote and highlight the advantages of one and two year contracts on the company's media accounts. Conduct periodic reviews for programs with month to month contacts
- Conduct periodic reviews for fiber optic internet service, TV and movie streaming , for example in terms of cost or poor quality
- Do more to post the importance of using online security and online backup to attract consumers to use these services.

Business Benefit

- PT X can predict consumers who have the potential to switch services to other companies so that PT X can make strategies to retain these consumers.
- The company gets an idea of what aspects need to be improved or become the main points to attract consumers
- PT X can save costs because retaining customers is easier than attracting new customers.

2. Business Background and Objective

Business Background

PT X is a company engaged in telecommunications. In the previous month there were consumers who left the company's services. If it continues, the company will lose consumers, while every company is competing to attract consumers. PT X is looking for the cheapest way so that consumers who have subscribed do not switch to services at other companies

Objective

Analyze the data and predict the churn of users (to identify people who will and will not renew their contract)

Data

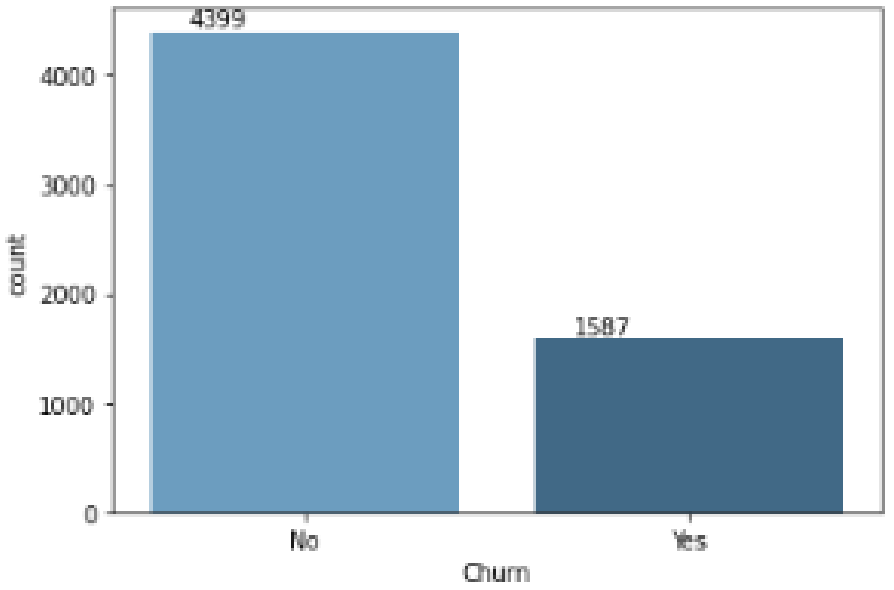
	Unnamed: 0	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProt
0	1869	7010-BRBUU	Male	0	Yes	Yes	72	Yes	Yes	No	No internet service	No internet service	No internet
1	4528	9688-YGXVR	Female	0	No	No	44	Yes	No	Fiber optic	No	Yes	
2	6344	9286-DOJGF	Female	1	Yes	No	38	Yes	Yes	Fiber optic	No	No	
3	6739	6994-KERXL	Male	0	No	No	4	Yes	No	DSL	No	No	
4	432	2181-UAESM	Male	0	No	No	2	Yes	No	DSL	Yes	No	

Shape of data: `(5986, 22)`

Columns name:

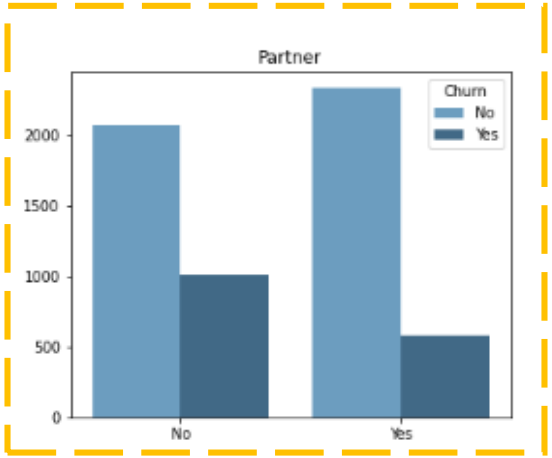
'Unnamed: 0', 'customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents', 'tenure', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges', 'TotalCharges', 'Churn'

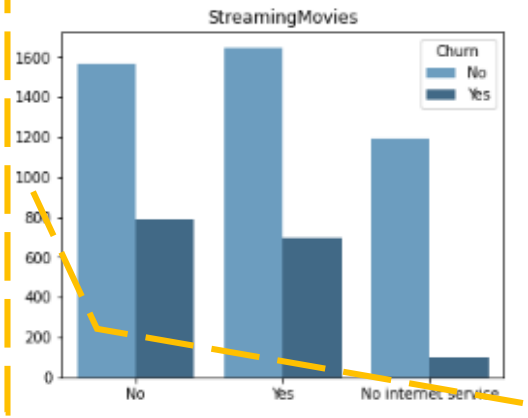
3. Exploratory Data Analysis

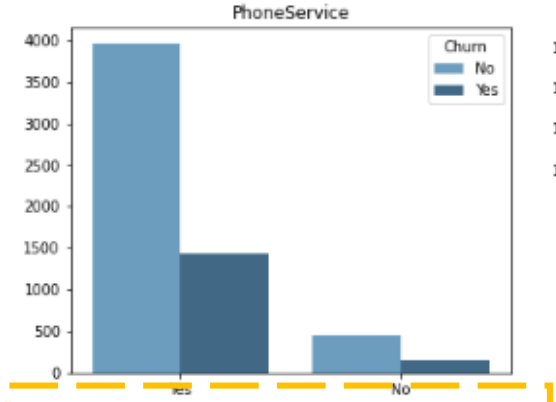


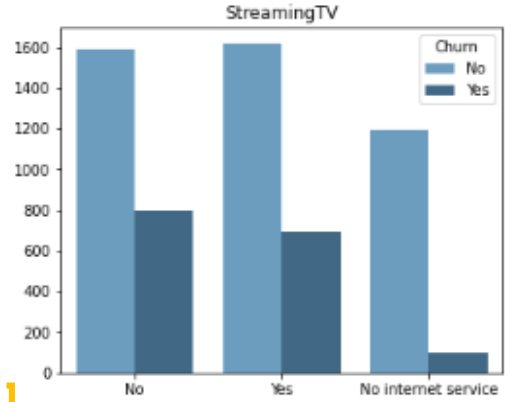
There is a data imbalance between churners and non-churners. On the data, non-churners far more churners. So that before making a model it is necessary to address the problem of data imbalance first. The data set contains information about Telco customers where each row represents a unique customers and the columns are information regarding customers' services. The column "Churn" indicate whether the customer left the company within the last month. There are a total of **5986** customers in the dataset among which **1587** left within the last month. With a churn rate that high, **26.51%**, Telco may run out of customers in the coming months if no action is taken.

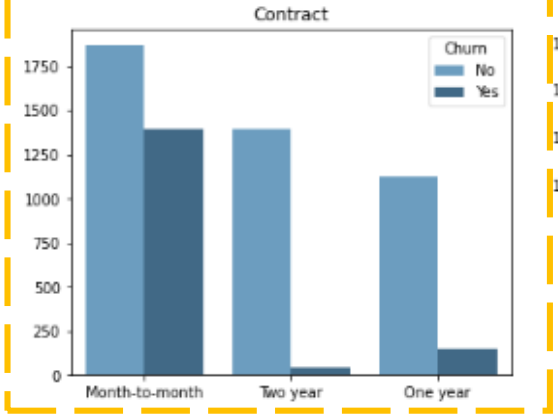
3. Exploratory Data Analysis

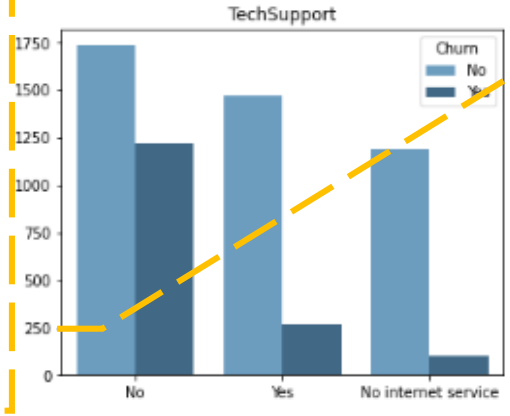








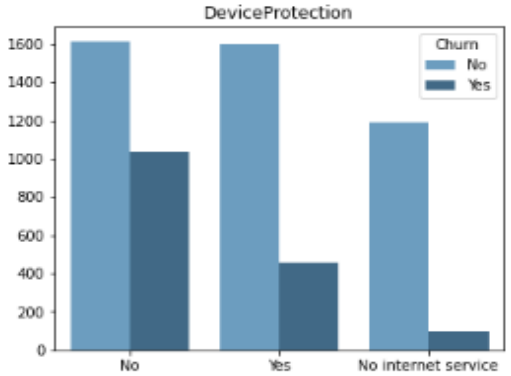


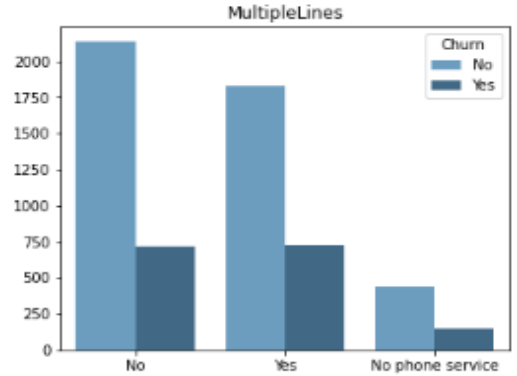


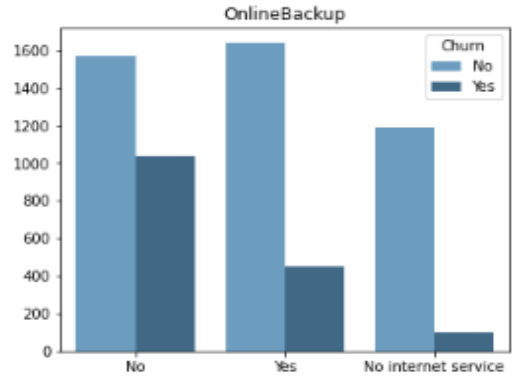
The churn rate of unmarried customers is higher than the churn rate of married customers, while the highest chance of loyal consumers lies with married consumers

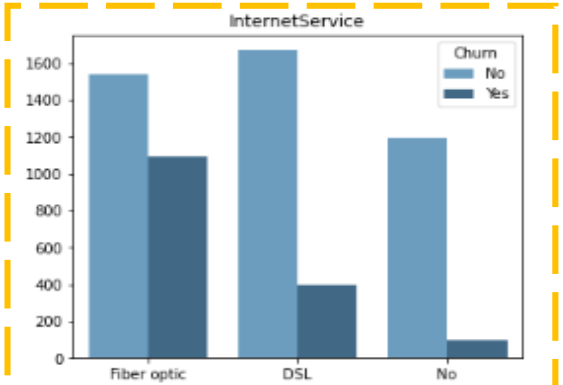
The churn rate for month-to-month contracts is higher than for other contract terms.

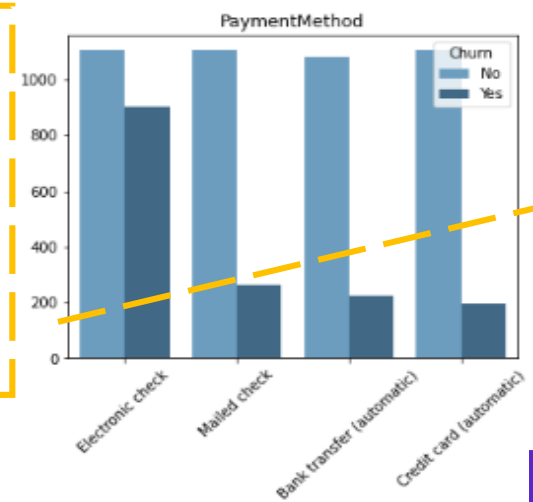
3. Exploratory Data Analysis





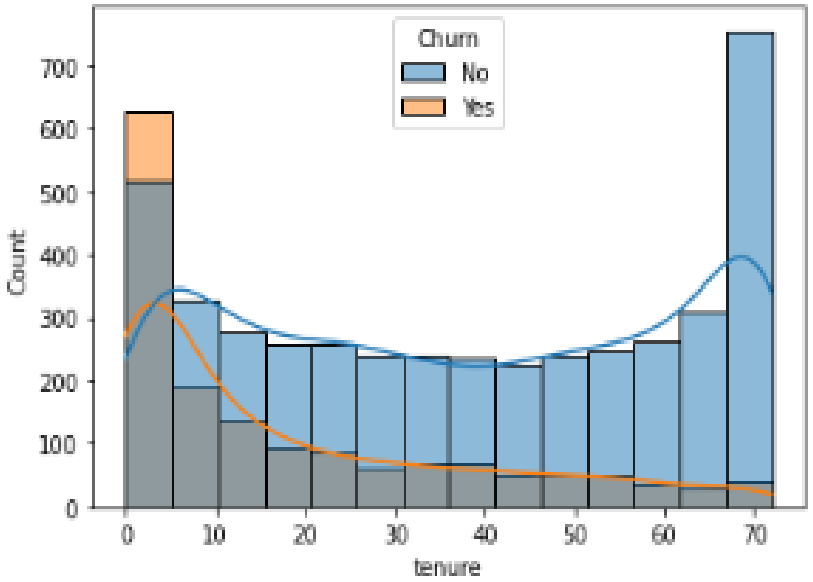


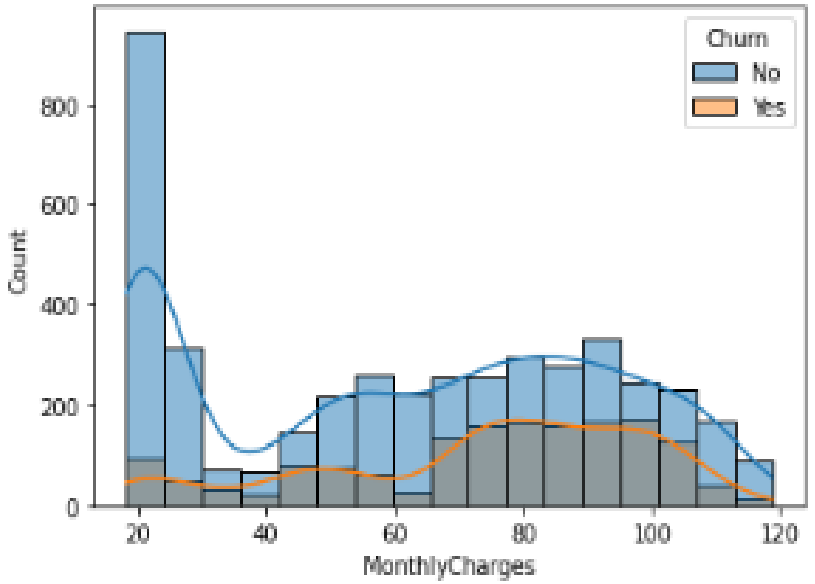




The highest chance of churn lies in internet service (fiber optic), while the highest chance of loyal customers lies in internet service (DSL).

3. Exploratory Data Analysis





Someone who has been a client of the company for **less than 5 months** has the potential to become an unfaithful customer. Customers who have subscribed for **more than 65 months** tend to be loyal

Disloyal customers tend to pay the more important fees, which are **between 65 and 105 dollar**. Loyal customers tend to pay more important fees ranging from **less than 30 dollar**

4. Data Preparation and Feature Engineering

Check number of missing value

```
Unnamed: 0          0
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges        0
Churn               0
```

1

2

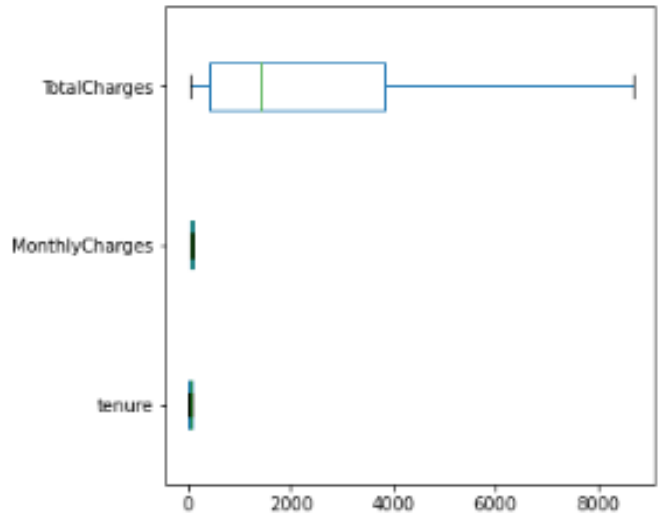
3

4

5

Cleaning Data

Check outliers using boxplot



Check for unique values

```
TotalCharges
            10
20.2        10
19.75        8
19.55        7
20.05        6
            ..
3815.4       1
259.65       1
6889.8       1
2435.15      1
3969.4       1
Name: TotalCharges, Length: 5611, dtype: int64
```

Strange value

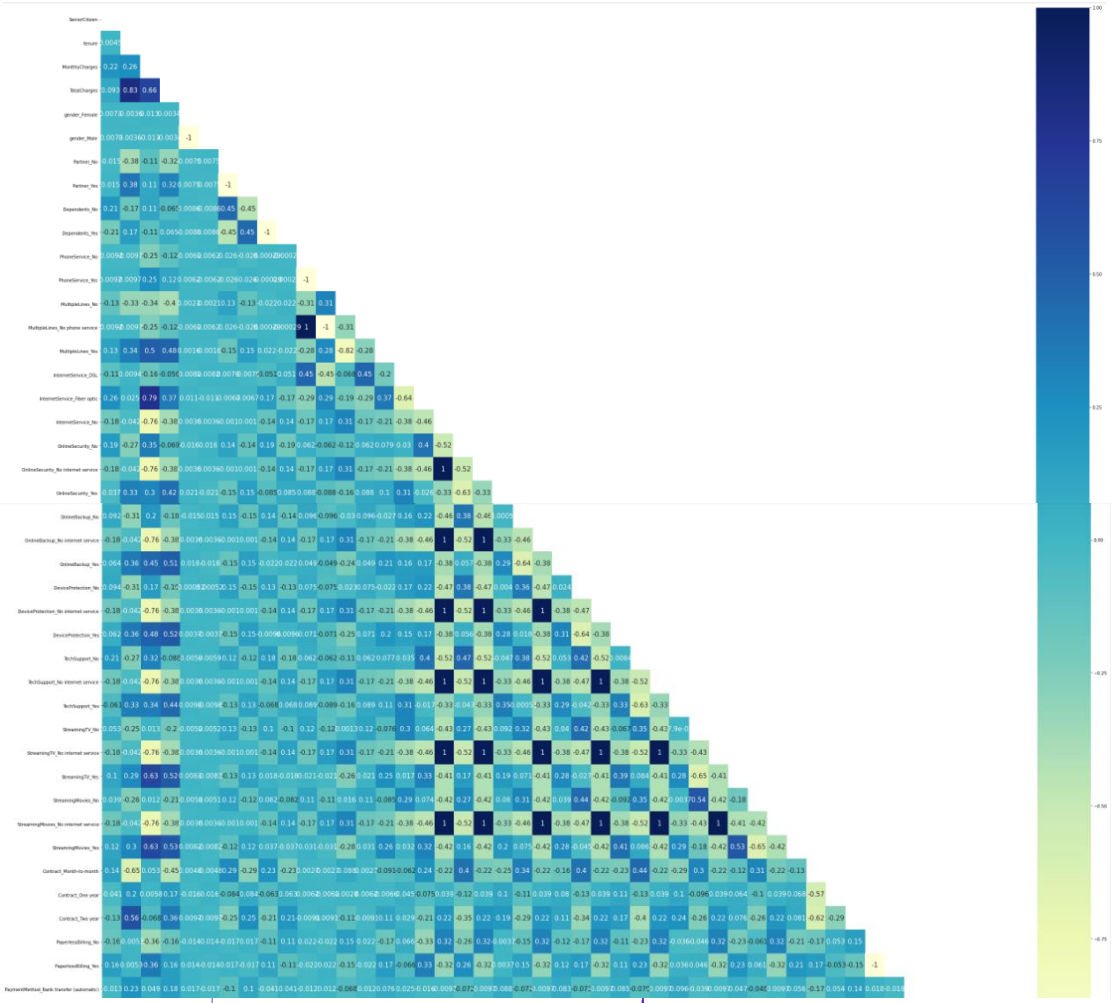
Remove ‘ ‘ value

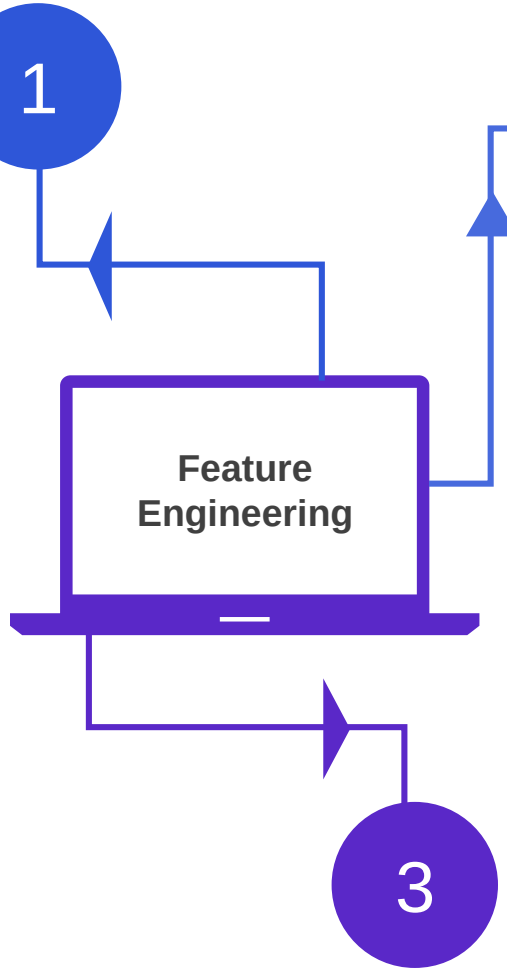
(5986, 22) → (5976, 22)

Remove costumer ID and convert all categorical variable to dummy variable

4. Data Preparation and Feature Engineering

Check correlation between feature





Remove some features that cause high correlation values (1)

Remove some features that cause very high correlation values (>0,8 or < -0,8)

5. Modeling and Evaluation

Scaling tenure and MonthlyCharges column

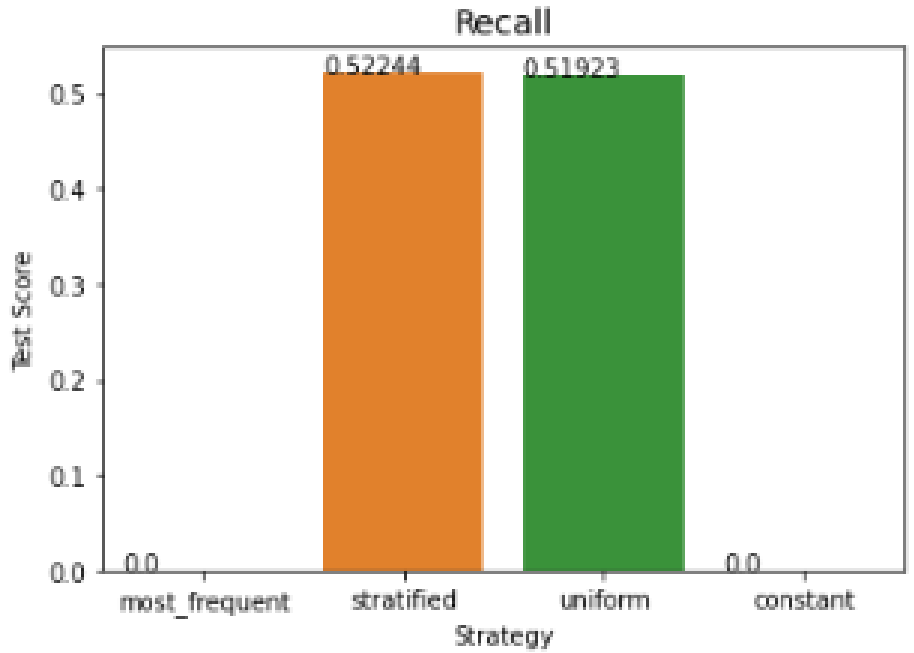
Define target and features

Split the data (train and test)

Overcoming the imbalance case (Churn column)

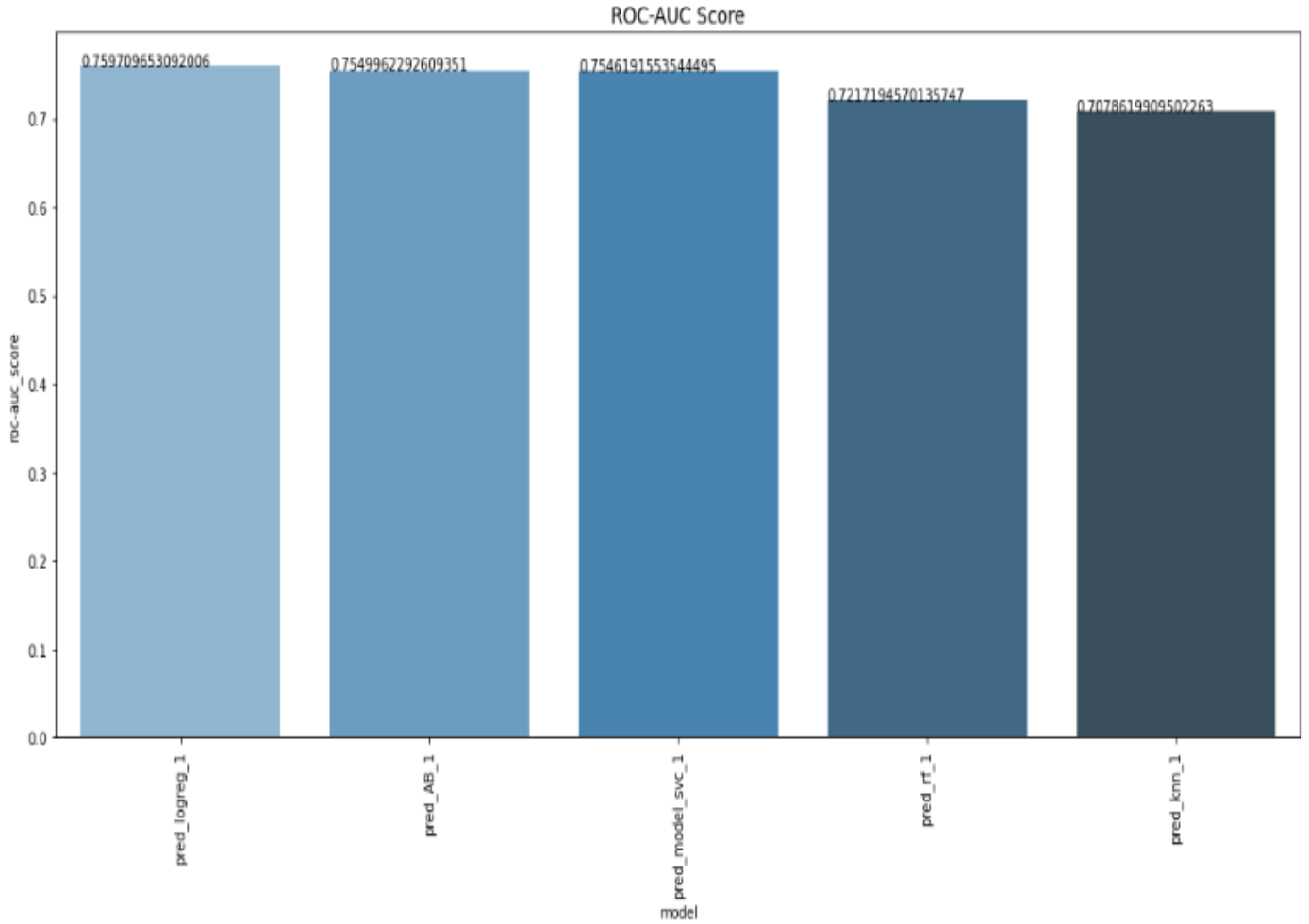
Baseline Model

Recall is more suitable for use in this case because it calculates the percentage of actual positives a model correctly identified (True Positive)



5. Modeling and Evaluation

Model	
KNN	Recall Train Data: 95.86% Recall Test Data: 75.64%
Logistic Regression	Recall Train Data: **79.57%** Recall Test Data: **80.45%**
Random Forest	Recall Train Data: 99.94% Recall Test Data: 59.62%
SVM	Recall Train Data: 88.25% Recall Test Data: 74.68%
AdaBoost	Recall Train Data: 85.79% Recall Test Data: 75.32%



The highest roc-auc score was found in the logistic regression model (75.97%)

5. Modeling and Evaluation

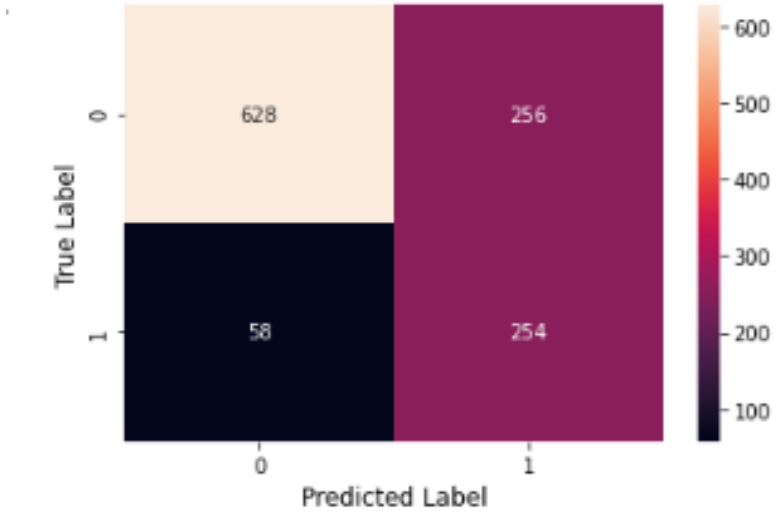
Hyperparameter Tuning

Logistic Regression

Recall Train Data: **80.37%**

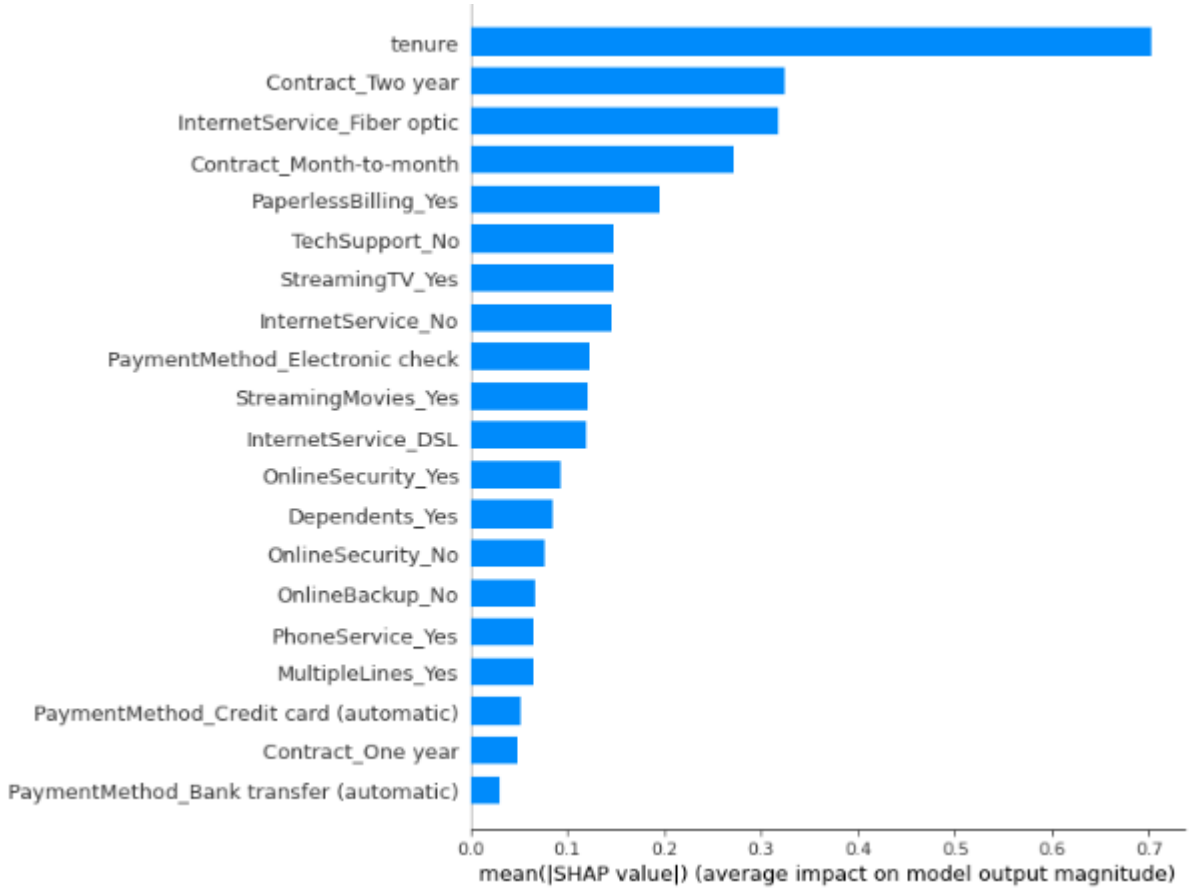
Recall Test Data : **81.41%**

ROC-AUC score : **76.23%**



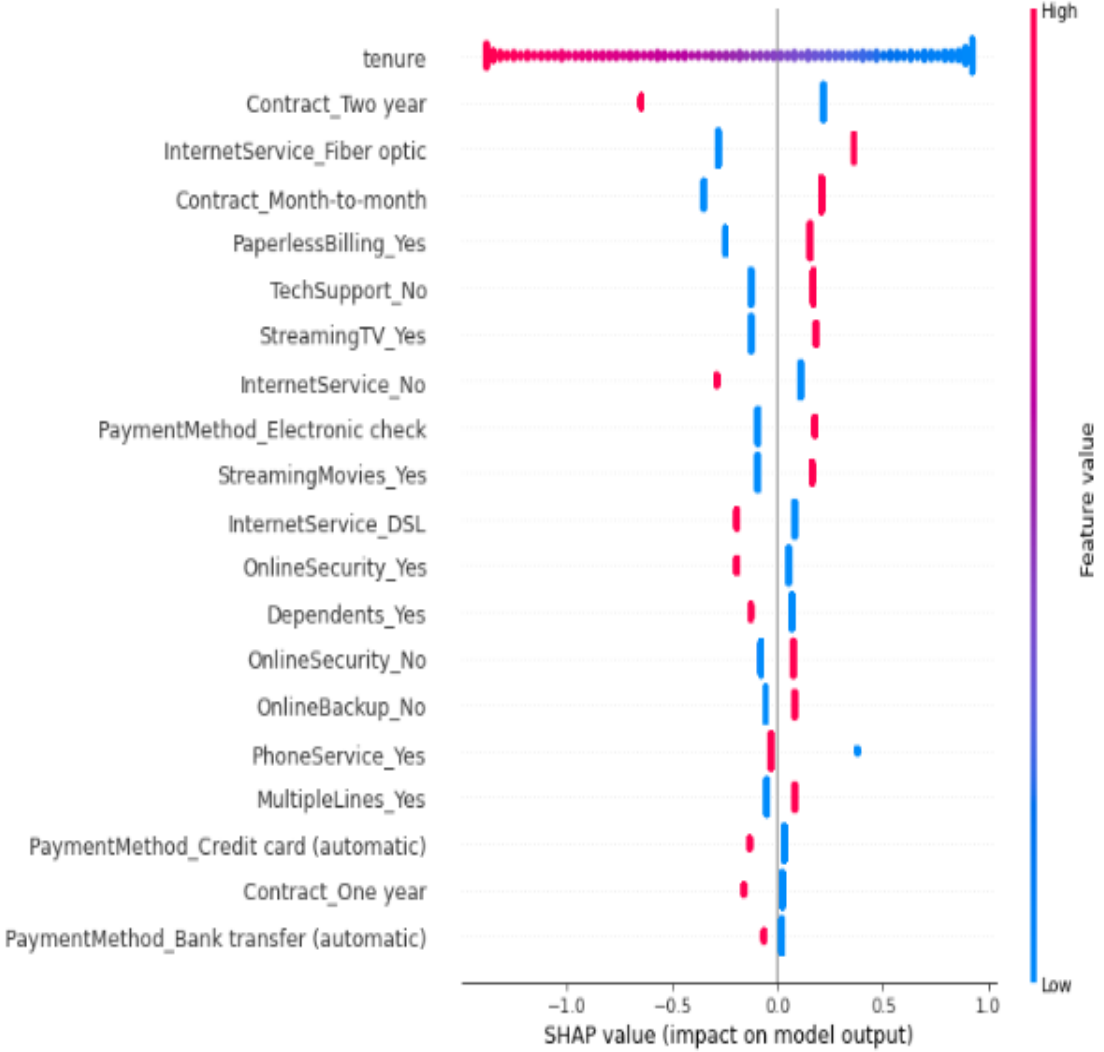
There are 254 disloyal consumers predicted correct

Feature Importance



The most important feature in the logistic regression model used in this case is tenure (the length of time consumers use the facility).

6. Recommendation



There needs to be more attention to new users, for example in the form of discounts or raffle points to make new users feel more comfortable using the service

promote and highlight the advantages of one and two year contracts on the company's media accounts. Conduct periodic reviews for programs with month to month contacts

Conduct periodic reviews for fiber optic internet service, for example in terms of price comparisons from competing companies or in terms of internet speed.

Do more to post the importance of using online security and online backup to attract consumers to use these services.

Periodic reviews are required for TV and movie streaming facilities, for example in terms of cost or poor quality

7. Simulation

The following is a simulation taken from the test data

	SeniorCitizen	gender_Male	Partner_Yes	Dependents_Yes	PhoneService_Yes	MultipleLines_Yes	InternetService_DSL	InternetService_Fiber optic	InternetService_No	OnlineSecurity_No	OnlineSecurity_Yes	Onli
5968	1	1	1	0	1	0	0	1	0	1	0	
4339	0	0	0	0	1	0	0	1	0	1	0	
2587	1	1	0	0	1	1	0	1	0	1	0	
4779	1	1	0	0	1	1	0	1	0	1	0	
2790	1	1	0	0	1	0	1	0	0	1	0	
...	...	...	...	...	...	...	...	...	...	...	...	
5480	0	0	0	0	1	1	0	1	0	1	0	
5952	0	0	1	1	1	0	0	1	0	1	0	
831	0	1	0	0	1	1	0	1	0	1	0	
2994	1	0	1	0	1	1	0	1	0	0	1	
1883	0	0	0	0	0	0	1	0	0	1	0	

510 rows × 31 columns

For example, a company will provide subsidies to consumers who have predicted that they will switch services to other companies by grouping consumers based on the 3 most important features and the amount of money consumers spend per month for the services used.

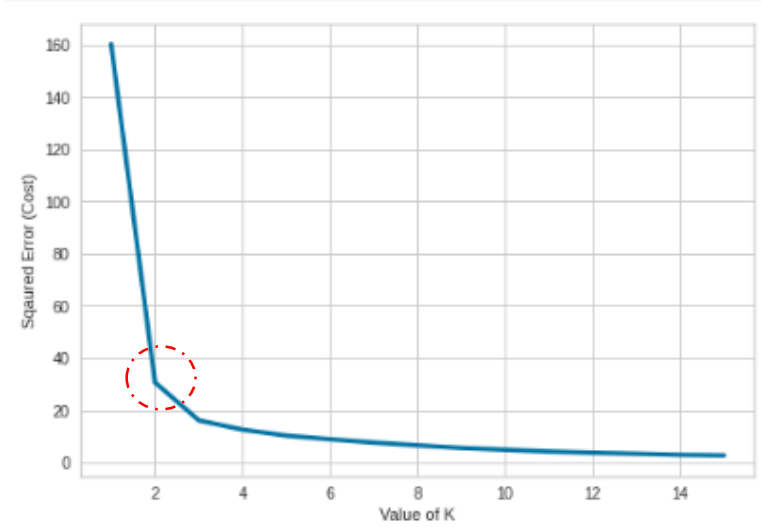
	tenure	MonthlyCharges	InternetService_Fiber optic	Contract_Two year
5968	0.366197	0.520398	1	0
4339	0.042254	0.518905	1	0
2587	0.084507	0.552239	1	0
4779	0.661972	0.803483	1	0
2790	0.014085	0.263184	0	0
...	...	...	...	...
5480	0.492958	0.861194	1	0
5952	0.000000	0.561692	1	0
831	0.028169	0.828856	1	0
2994	0.591549	0.758706	1	0
1883	0.000000	0.116418	0	0

510 rows × 4 columns

7. Simulation

K-Mean

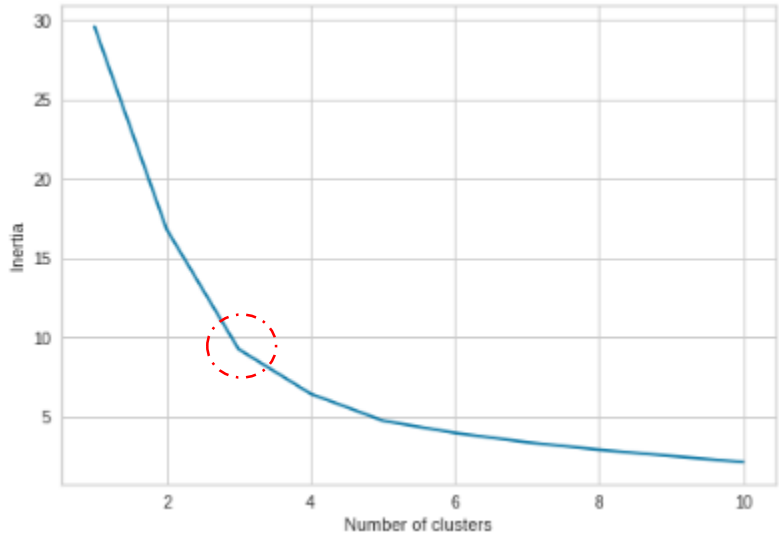
Elbow Method



The optimum k value is 2

Cosine K-Mean

Elbow Method



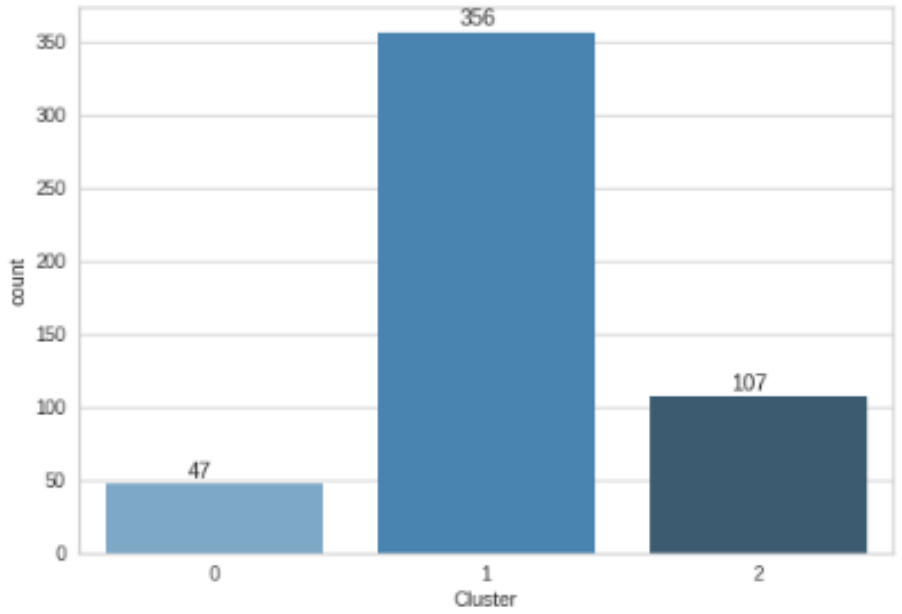
The optimum k value is 3

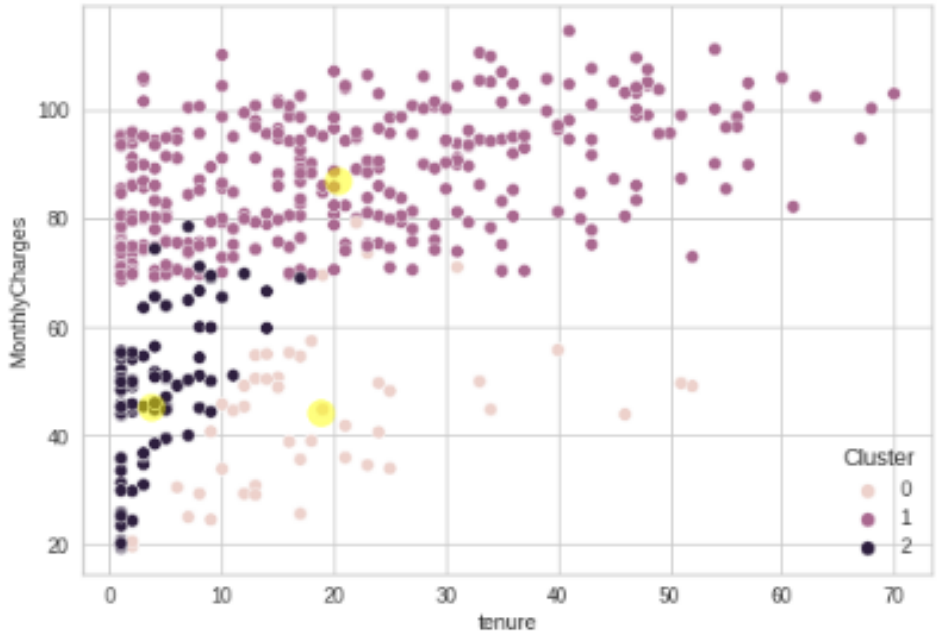
Evaluation

Kmeans : 74.54%

Cosine kmeans : **87.90%**

7. Simulation





Group 1 (0) --> the majority group consists of consumers who have been using the service for a long time and the amount of money spent is classified as lower middle class

Group 2 (1) --> the majority group consists of consumers who spend money on services with medium to high costs

Group 3 (2) --> the majority group consists of consumers who have just used the service (new users) with the amount of money spent belonging to the middle to lower class

7. Simulation

With this model, companies can maximize their efforts to retain consumers. For the example:

Since there are three groups of consumers who are predicted to leave the company's services, there are three solutions or to overcome this. Suppose each cluster is given a different subsidy (each person's cost).

Cluster 1= $10, Cluster 2= $25, and Cluster 3= $15

- Group 1 (0) --> Subsidies are given for DSL internet services or phone service or discount for 2 years contract
- Group 2 (1) --> Subsidies are given for fiber optic internet services or discount for 2 years contract
- Group 3 (2) --> Subsidies are given for DSL internet services or phone service or discount for 2 years contract

The subsidy for group 3 is bigger than group 1 because consumers in group 3 tend to be consumers who are new to using the service, so they need to get more attention in terms of subsidies in order to feel comfortable using the service. While group 1 is a collection of consumers who have been using the service for a long time, so that it places more emphasis on improving service quality

Budget **$10975**

7. Simulation

In addition, companies need to provide subsidies or discounts to consumers who are predicted to be loyal. It aims to increase the level of consumer loyalty. Suppose each consumer is given a subsidy or discount of 5 dollars

Budget **$3430**

Total Budget **$14405**

Note: there are several possibilities that will happen.

1.consumers are interested in discounts or subsidies
2.consumers are not interested in discounts or subsidies but still use the service
3.consumers are not interested in discounts or subsidies and switch services

7. Simulation

If without using a model that can predict churn cases, the company will incur different costs. This is because the company has to find new consumers due to the loss of 510 consumers. Here's the explanation.

TV ad cost: $115.000 (30-second)
TV ad production: $2.000 (minimum cost)
Total : $115.000 + $2.000 = **$117.000**

Note: there are several possibilities that will happen.

1.Can attract new customers as many as or more than customers who switch services
2.Unable to attract as many new customers as the number of consumers who moved


THANK YOU