IDS 561: ANALYTICS FOR BIG DATA

GROUP 9 - PROJECT REPORT

TELECOMMUNICATION CHURN PREDICTION

Sr. No.	Name	UIN
1	ADITYA NAIGAONKAR	671279933
2	JITESH PATIL	653037840
3	NIRMAL TANJORKAR	665035319
4	VIDHIT TAMBE	651729186

PROBLEM STATEMENT:

One of the most pressing problems the telecoms sector is dealing with is customer turnover. Compared to keeping an existing customer, acquiring a new one is more expensive. As a result, the telecommunications industry's primary method for preventing customer churn is customer churn prediction. Predicting customer churn aims to locate departing clients for a telecommunications service provider in advance. The telecom service provider could plan their client retention strategy with the help of a customer churn forecast. With the application of data mining techniques, the industry's large volume of data becomes the key resource for forecasting customer attrition.

Nowadays, corporations use sophisticated algorithms to anticipate which consumers are most likely to leave the company. Companies can develop customer retention strategies to reduce potential losses by employing such algorithms to identify in advance the clients who are most likely to stop using their services. Our attempt in this project is to build a machine-learning model to predict customer churn using different classification algorithms.

STATISTICS:

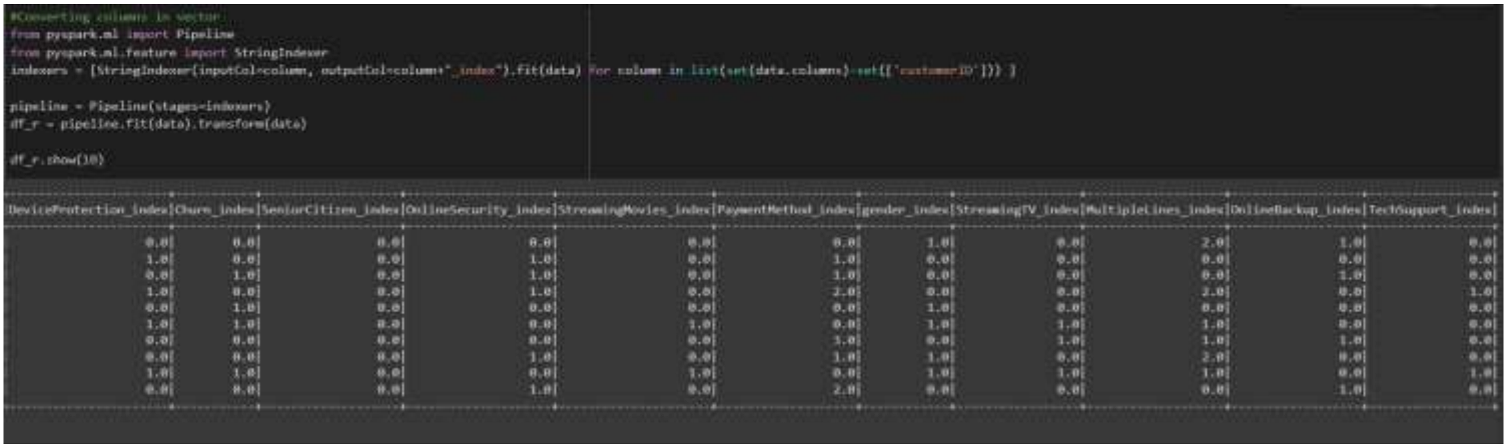
- Average annual turnover rate: 10% to 67%
- If they had better plans, 40% of the customers wouldn't have changed.
- 79% of customers confessed venting to others about how unhappy they were with the telco's level of service.

DATASET:

The Telco Customer Churn dataset is what we are using which is available on the Kaggle website and has 7043 rows; each row corresponds to a customer. The following are the qualities or characteristics:

- **Customer ID**: Each customer is given a different ID.
- **Gender:** Identifies the client's gender. Male and female are the two sorts.
- **Senior Citizen (0/1):** Indicates whether the customer is a senior. 1 means "yes," while 0 means "no."
- **Partner (Yes/No):** This field indicates whether or not the customer has a partner.
- **Dependents (Yes/No):** This field lets you know whether the consumer has any dependents.
- **Tenure:** This number represents how many months the customer has been a customer of the telecom business.
- **Phone Service:** This field lets you know whether the client has residential phone service.

- **Multiple lines:** If the consumer has multiple residential lines, it will say "Multiple Lines." There are three categories: "No phone service," "No," and "Yes."
- **Internet Service:** Describes the customer's type of internet service. DSL, fiber optics, and no are the three types.
- **Online Security (Yes/No):** The customer's level of internet security is indicated by a "Yes" or "No" response under Online Security.
- **Online Backup (Yes/No):** Indicates whether or not the customer has an online backup (Yes/No).
- **Device Protection (Yes/No):** The customer's device protection status is indicated by a "Yes" or "No" response for device protection.
- **Tech Support (Yes/No):** Indicates whether the consumer has access to tech support (Yes/No).
- **Streaming TV (Yes/No):** Whether the consumer has access to a TV streaming capability is indicated by the "Streaming TV" (Yes/No) field.
- **Streaming Movies (Yes/No):** Indicates whether the customer has access to a feature that allows them to stream movies (Yes/No).
- **Contract:** Describes the type of agreement the consumer has with the telecom provider. There are three types: "Month-to-Month," "One Year," and "Two Year."
- **Paperless Billing (Yes/No):** Whether the consumer has chosen the paperless billing option is indicated by the "Paperless Billing" (Yes/No) field.
- **Payment Method:** The sort of payment method the consumer has selected is indicated by the payment method field. Electronic checks, mailed checks, bank transfers, and credit cards are the available types.
- **Monthly Charges:** A continuous variable, it is. shows the monthly cost that the consumer pays to the telecom service provider.
- **Total Charges:** This variable is continuous. shows the whole cost that the consumer has paid to the telecom service provider.
- **Churn (Yes/No):** This value indicates if a client has stopped using the telecom service provider.



```
#Converting columns in vector
from pyspark.ml import Pipeline
from pyspark.ml.feature import StringIndexer
indexers = [StringIndexer(inputCol=column, outputCol=column+"_index").fit(data) for column in list(set(data.columns)-set(['customerID'])) ]

pipeline = Pipeline(stages=indexers)
df_r = pipeline.fit(data).transform(data)

df_r.show(10)
```

```
+----------------------+-----------+-------------------+--------------------+---------------------+-------------------+------------+-----------------+-------------------+------------------+-----------------+
|DeviceProtection_index|Churn_index|SeniorCitizen_index|OnlineSecurity_index|StreamingMovies_index|PaymentMethod_index|gender_index|StreamingTV_index|MultipleLines_index|OnlineBackup_index|TechSupport_index|
+----------------------+-----------+-------------------+--------------------+---------------------+-------------------+------------+-----------------+-------------------+------------------+-----------------+
|                   0.0|        0.0|                0.0|                 0.0|                  0.0|                0.0|         1.0|              0.0|                2.0|               1.0|              0.0|
|                   1.0|        0.0|                0.0|                 1.0|                  0.0|                1.0|         0.0|              0.0|                0.0|               0.0|              0.0|
|                   0.0|        1.0|                0.0|                 1.0|                  0.0|                1.0|         0.0|              0.0|                0.0|               1.0|              0.0|
|                   1.0|        0.0|                0.0|                 1.0|                  0.0|                2.0|         0.0|              0.0|                2.0|               0.0|              1.0|
|                   0.0|        1.0|                0.0|                 0.0|                  0.0|                0.0|         1.0|              0.0|                0.0|               0.0|              0.0|
|                   1.0|        1.0|                0.0|                 0.0|                  1.0|                0.0|         1.0|              1.0|                1.0|               0.0|              0.0|
|                   0.0|        0.0|                0.0|                 0.0|                  0.0|                1.0|         0.0|              1.0|                1.0|               1.0|              0.0|
|                   0.0|        0.0|                0.0|                 1.0|                  0.0|                1.0|         1.0|              0.0|                2.0|               0.0|              0.0|
|                   1.0|        1.0|                0.0|                 0.0|                  1.0|                0.0|         1.0|              1.0|                1.0|               0.0|              1.0|
|                   0.0|        0.0|                0.0|                 1.0|                  0.0|                2.0|         0.0|              0.0|                0.0|               1.0|              0.0|
+----------------------+-----------+-------------------+--------------------+---------------------+-------------------+------------+-----------------+-------------------+------------------+-----------------+
```

Telco churn dataset (before performing EDA)

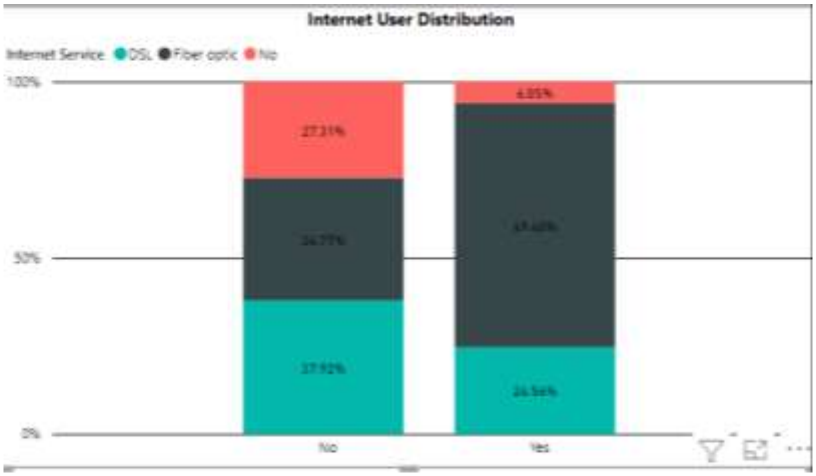
EXPLORATORY DATA ANALYSIS:

Following are some key findings after exploratory data analysis and removing unnecessary (unwanted or duplicate) variables.

- The dataset has an excel tabular format.
- Each customer has 21 attributes (columns).
- The churn (Yes/No) is the Target variable in our situation.
- 1869 records out of the total of 7043 correspond to churn (yes), and 5174 records do as well (No).
- There are only 2 continuous characteristics. The remaining 19 are classified

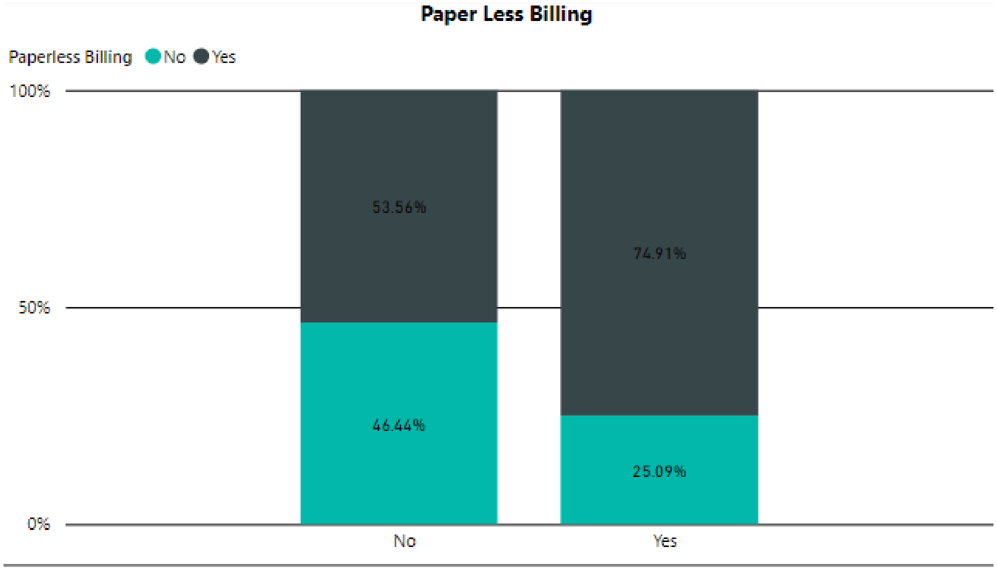
VISUALIZATION USING POWER BI:

Continuing with the EDA, we analyzed the data using an interactive Power BI dashboard and visualization to gain further insights into the dataset before building our model. Below are the examples of visualization we plotted.



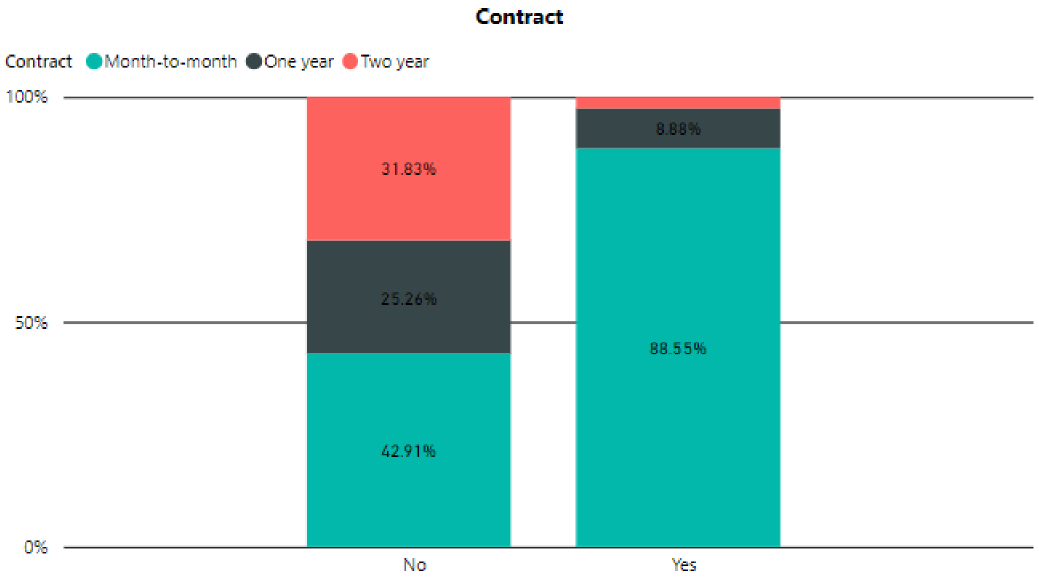
Internet user distribution

More % of Customers choosing fiber optics: The plot shows that the churn rate drops when a client chooses fiber optic service. When a consumer chooses DSL, we see a similar pattern.



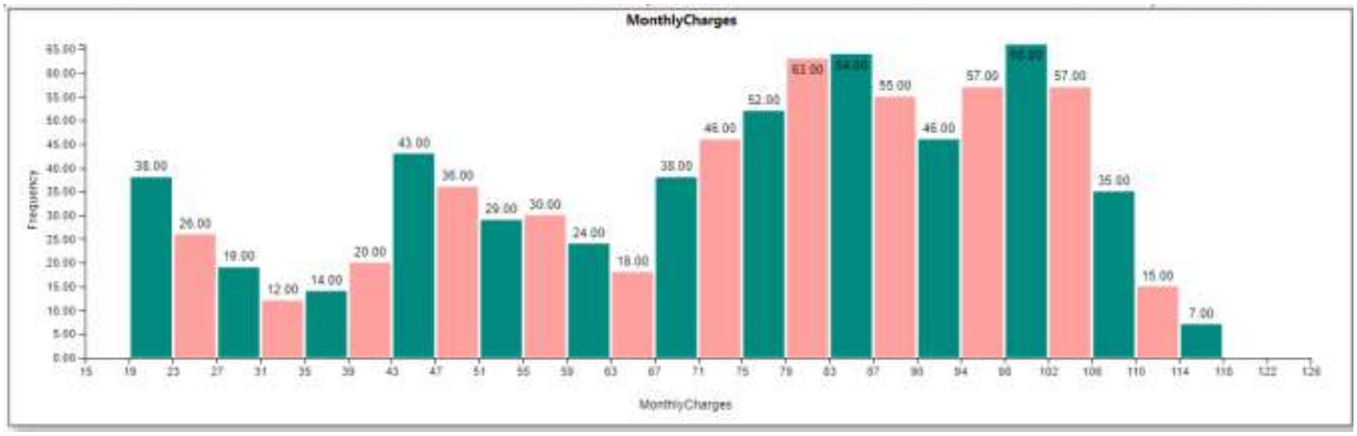
paperless billing

The churn rate drops when a customer chooses a paperless billing method

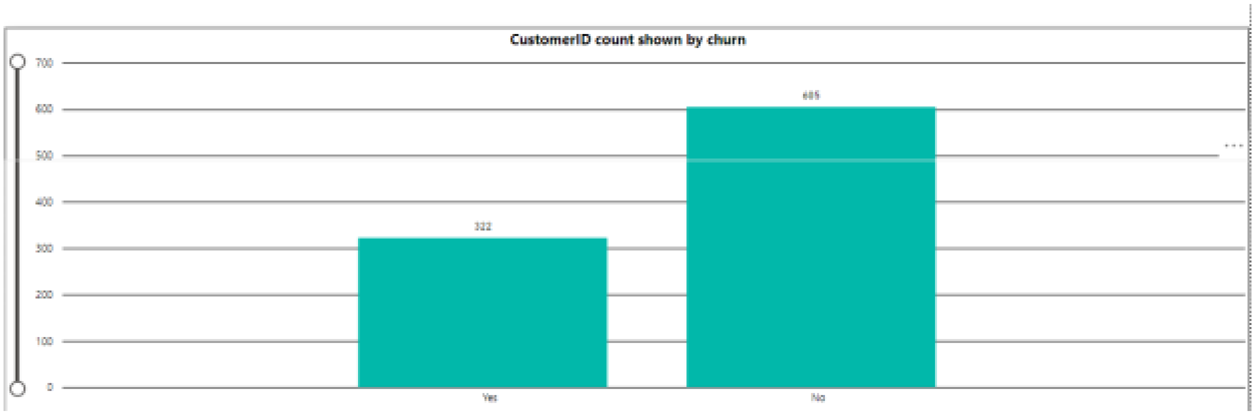


Churn rate by contract

As we can see, the churn rate is reduced by ~90% when the company offers a two-year contract. Like the two-year contract.



This graph shows a histogram of monthly charges, and we can see that consumer spending on monthly bills has a slight rightward skewness.



Total Churn rate by Customer Id

MODEL BUILDING:

Based on our dataset, we use Python to develop machine-learning models to forecast customer attrition. For data preprocessing and model construction, we used Pyspark.
Below the key findings and steps taken to build our model are listed: -

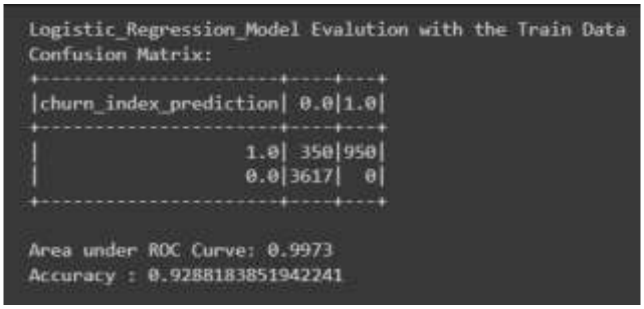
- The data is read into a Pyspark data frame in multiline format after starting a spark session. The label (string) indexer maps the columns to label indices, followed by pipelining and encoding.
- Utilizing the vector assembler, the categorical data is encoded and vectorized. The vectorized features, their labeled index, and the encoded target variable (customer churn) make up the final data frame that is used to build and test models. Customer churn "Yes" and "No" is encoded as "1" and "0," respectively.
- The train-to-test ratio for the dataset is divided into 70:30. **Logistic regression** and **naive Bayes** are the machine learning methods utilized in the model construction process.
- A regularization parameter is employed to prevent overfitting of the data in the logistic regression model. The regularization parameter with the lowest hold-out cross-validation error is determined to be the optimal regularization parameter.
- Laplace smoothing and multinomial classification are both employed in the naive Bayes model. Based on test data, the models developed generate predictions.
- Through metrics like the confusion matrix, accuracy, and the area under the true positive vs. false positive ROC curves, these predictions are used to assess models. These ROC curves allow for the selection of appropriate categorization threshold values based on various scenarios.

NAIVE BAYES: A group of supervised learning algorithms known as naive Bayes techniques utilizes Bayes' theorem with the "naive" assumption that each pair of features is conditionally independent given the value of the class variable. A Naive Bayes classifier, to put it simply, believes that the presence of one feature in a class has nothing to do with the presence of any other feature. Simple to construct and especially helpful for very big data sets is the naive Bayes model. Along with being straightforward, Naive Bayes is known to perform better than even the most complex classification techniques.

LOGISTIC REGRESSION: This kind of statistical analysis sometimes referred to as a logit model, has several uses in machine learning applications as well as predictive modeling. In this analytics approach, the dependent variable is finite or categorical: either A or B (binary regression) or a range of finite options A, B, C, or D (multinomial regression). By estimating probabilities using a logistic regression equation, it is possible to analyze the relationship between the dependent variable and one or more independent variables using statistical software. This kind of analysis can assist you in estimating the probability of an event occurring or a decision being made.

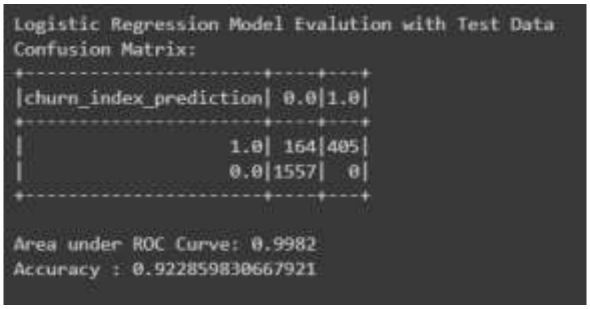
FINAL RESULTS BASED ON PREDICTION MODELS:

- The logistic regression model has a similar area under the true positive vs. false positive ROC curve of ~99.00 % and a training and testing accuracy of ~93%.



```
Logistic_Regression_Model Evalution with the Train Data
Confusion Matrix:
+----------------------+----+---+
|churn_index_prediction| 0.0|1.0|
+----------------------+----+---+
|                   1.0| 350|950|
|                   0.0|3617|  0|
+----------------------+----+---+

Area under ROC Curve: 0.9973
Accuracy : 0.9288183851942241
```

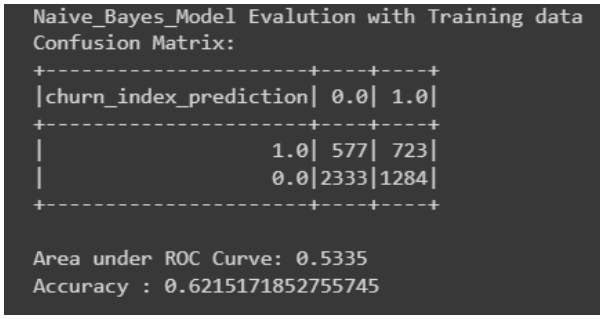


```
Logistic Regression Model Evalution with Test Data
Confusion Matrix:
+----------------------+----+---+
|churn_index_prediction| 0.0|1.0|
+----------------------+----+---+
|                   1.0| 164|405|
|                   0.0|1557|  0|
+----------------------+----+---+

Area under ROC Curve: 0.9982
Accuracy : 0.922859830667921
```

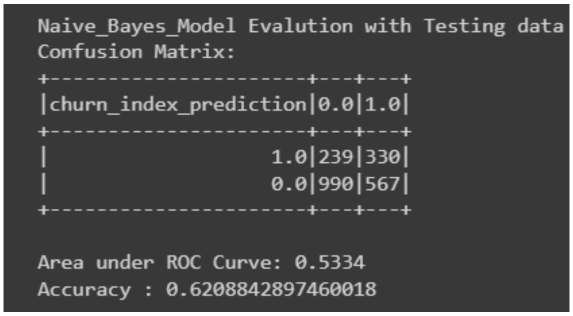
The regularization parameter in use is to blame for this. A good model fit is implied by the same train/test accuracies and AUC.

- The multinomial naive Bayes model yields a training and testing accuracy of ~62%; the similarity between the two values suggests a strong model fit. ROC curves have training and testing AUCs of ~53% each.



```
Naive_Bayes_Model Evalution with Training data
Confusion Matrix:
+----------------------+----+----+
|churn_index_prediction| 0.0| 1.0|
+----------------------+----+----+
|                   1.0| 577| 723|
|                   0.0|2333|1284|
+----------------------+----+----+

Area under ROC Curve: 0.5335
Accuracy : 0.6215171852755745
```



```
Naive_Bayes_Model Evalution with Testing data
Confusion Matrix:
+----------------------+---+---+
|churn_index_prediction|0.0|1.0|
+----------------------+---+---+
|                   1.0|239|330|
|                   0.0|990|567|
+----------------------+---+---+

Area under ROC Curve: 0.5334
Accuracy : 0.6208842897460018
```

FINAL RESULTS BASED ON DATA VISUALIZATIONS:

- By offering customers a long-term contract of 2 years, the customer churn rate can be reduced.
- People with very high or very low tenure and senior citizens are leaving the service. The company should introduce new plans and services to senior citizens to reduce the churn rate.
- There is no problem with monthly pricing or the manner of payment, but most clients cannot afford a two-year subscription, so the business needs to come up with new offers and schemes.