TELECOM CUSTOMER CHURN

Student id – 21133614 **Name – Nirvan Kamble**

1. Domain information.

The telecommunications industry has grown to be one of the most important in developed countries. The level of competition has risen because of technological advancements and an increase in the number of operators. Companies are working hard to stay afloat in this competitive market, employing a variety of strategies. To increase revenues, three main approaches have been suggested.

Three main approaches which as follows:

- **Acquire new customers.**
- **Upsell existing customers.**
- **Increase customer retention.**

However, comparing these strategies based on the value of return on investment (RoI) has revealed that the third strategy is the most profitable, demonstrating that retaining an existing customer costs much less than acquiring a new one in addition to being considered much easier than upselling strategy.

To apply the third, companies must reduce the possibility of customer churn, also known as "the customer movement from one provider to another". In a business context, customer attrition purely leads to consumers switching from one business service to the another. Customer churn, also recognised as subscriber churn, is a term that refers to the process of customers switching from one service provider to another completely anonymous Customer churn is a major concern in service sectors with high levels of competition.

Predicting which customers are likely to leave the company, on the other hand, offers a relatively large additional revenue source if done early on. Many studies have also shown that **machine learning** technology is extremely effective at predicting this situation. This technique is used by learning from previous data. From the standpoint of machine learning, churn prediction is a supervised (i.e., labelled) problem defined as follows: Given a predefined forecast horizon, the goal is to predict future churners over that horizon using data associated with each network subscriber. Churn Prediction is a phenomenon that is used to identify potential churners before they leave the network. This assists the CRM department in preventing subscribers who are likely to churn in the future by implementing the necessary retention measures.

Customer Churn –

The term 'churn' is formed from the words change and turn. It denotes the termination of a contract. Churn can be classified into three types:

- **active/deliberate** - the consumer chooses to cancel his contract and switch to a different supplier. Reasons for this may include dissatisfaction with service quality (e.g., failure to meet service level agreements), excessive costs, high level of competition price plans, no rewards for customer loyalty, a lack of understanding of the service scheme, poor support, a lack of information about the causes and anticipated response times for service problems, a lack of consistency or fault resolution, privacy concerns, and so on.
- **rotational/incidental** - the consumer terminates the contract without intending to switch to a competitor. Reasons for this include changes in the environment that prevent the customer from requiring the service in the future, such as financial difficulties that make payment impossible; or a change in the customer's geographical location to a location in which the business is not present or the service is unavailable.
- **Passive / non-voluntary** - the contract is terminated by the corporation.

Customer churn can occur because of any factor, and it is hard to evaluate with so much data to consider. It is critical to characterize which potential churners are of further interest to the company, for example, which customers are effective in creating more revenue (these are typically customers who generated significant revenue and then found a good offer with a decent loyalty program at a competitor), and which customers are not interested, for example, because they are identified as risky. The marketing department of the organization can then consider direct marketing tactics to retain critical consumers. Although churn is an inescapable issue, it may be managed and potential costs to the business reduced. This goal is supported through timely detection of potential churners, as well as effective retention measures.

Although churn is an inescapable event, it may be handled and the potential damage to the business reduced. This goal is supported through early detection of potential churners, as well as effective retention measures.

As a Machine Learning engineer, we can detect customer churn early and provide a proper answer to the organisation on which customers to focus on and which customers to ignore.

2. Problem Definition.

Wire line companies are facing greater competition from cable operators, as well as an increasing risk of disruption from OTT players. One or more of these powerful trends is putting pressure on telecom companies to respond with more competition, bundles, and price cuts. Given the industry's demanding changes, maintaining a customer base to minimise churn should be one of the top priorities for any high-ranking telecom executive. Furthermore, work with telecom companies all over the world has revealed that companies that implement an extensive, analytics-based methodology to base management can reduce churn by up to fifteen percent.

The following statement of the problem will be analysed, and a solution will be presented to a telecommunications firm to minimize churn and maximize the company's profit margin.

1. Develop a predictive model to classify customer churn risk.
2. Develop a predictive Model to see which Payment method will be used.
3. Suggest potential approaches to reduce customer churn.

Develop a predictive model to classify customer churn risk.

- In this problem statement, we will create a model using a machine-learning algorithm to predict whether a future customer will churn or not so that the company can take the necessary steps to retain the customer.

Develop a predictive Model to see which Payment method will be used.

- In this problem statement, we will assess the most used payment method by the customer, including how much churn occurs from that specific payment, and we will develop a model using a machine-learning algorithm to predict which payment method future consumers will use. This allows a company to target the most popular payment method group and offer a discount to retain customers for a longer period.

Suggest potential approaches to reduce customer churn.

- In this problem statement, we will provide a solution to the firm in terms as to which group of customers should receive the most discount and which customers should receive the least discount for future retention, so that the investment to reduce customer churn is properly invested and has a higher chance of return.

Data set description: -

As we all know, the data set serves as the beginning point for everything; it should contain sufficient data to allow the machine to learn about the problem. Scrap data available on the internet can be used to generate or develop datasets. In our case, we used data from the internet that was freely available on the Kaggle website.

Kaggle dataset link https://www.kaggle.com/blastchar/telco-customer-churn

A relevant data set is required to grasp why, and which customers are churning most and, further because the factors that are driving customers to change providers. the corporate has provided us with a dataset containing 7043 individuals with various attributes and telecom service preferences.

In the dataset, 21 variables might affect churning within the company. Customers who chosen to go away within the last month are mentioned as churn, and this is often our target variable. telephone service, multiple lines, internet, online security, online backup, device protection, tech support, TV streaming, and movie streaming are all services that customers have signed up for. Tenure, contract type, payment method, paperless billing, monthly charges, and total charges are all samples of customer account information. Customers' demographic information, like gender, age, and whether or not they have partners and youngsters. This dataset contains numerical values of the integer and float types, as well as categorical values of the string or object type.

Numerical columns - Monthly charges, Total charges, Tenure, Senior Citizen.

Categorical columns - Phone service, Multiple lines, Internet, Online security, Online backup, Device protection, Tech support, TV streaming, Movie streaming, Contract type, Payment method, Paperless billing, Dependents.

Let us see what each column means: -

Gender: Gender of the customer: Male and female

Senior Citizenship: Indicates whether or not the customer is 65 or older: Yes or No

Partner: Indicate whether or not the customer has a partner: Yes or No

Dependents: Indicates whether the customer has any dependents: Yes and no. Dependents can include children, parents, grandparents, and others.

Tenure Months: Indicates the total number of months the customer has been with the company as of the end of the specified quarter.

Phone Service: Indicates whether the customer has a home utility subscription with the company: Yes or No

Multiple Lines: Indicates whether or not the customer has multiple telephone lines with the company: Yes or No

Internet Service: Indicates whether the customer has Internet service with the company: DSL, Fibre Optic, and Cable are the only options.

Online Security: Indicates whether or not the customer is secure online.

In this dataset some columns as the most importance and some are not that useful for the predication just like there is no correlation between the ratio of churning customers and gender. Aside from gender, customer ID, senior citizen, and partner have no impact on churning. But on the other side, one of the most exciting data results is Internet service and dependents. If a customer has a dependent, such as children, those who would be less likely to churn. The internet is also important. Tenure seems to be another captivating factor that influences churning. One exciting feature of payment methods from there we can see which payment mode is mostly used and we can do some great analysis with it. Additional consideration, some variables such as monthly charges and multiple lines may appear important in terms of understanding customer churn. Tenure, fibre optic, contract, and special features are most valuable churning factors.

There is some limitation in this dataset also, there are two types of churn: involuntary churn and voluntary churn. The information supplied by telecom companies just concentrates on voluntary churn; however, involuntary churn has a significant impact on a company's revenue and customer service. Involuntary churn can be caused by natural causes such as death, but it can also be caused by problems such as card expiration, bounced checks, exceeding credit limits, and fraud (Campbell, 2019). As a result, the analysis is also constrained.

The most critical and crucial things to consider when predicting are false positives and false negatives. In the telecom industry, a false positive occurs whenever the predicted number of customers with no churn actually churns. If the prediction is yes to churn but the outcome is no churn, this is referred to as a false negative. In this case, false positives are critical. A false negative can result in an unexpected drop in revenue and brand reputation. It becomes much more expensive for a company to retain customers who haven't churned because in order to accelerate their growth and keep their revenue high, they must seek out new customers. Obtaining new customers and retaining existing customers incurs additional costs and time in marketing, advertising, and other facilities.

Dataset Exploration: -

Data exploration is the starting point of information in which data analysts use visualisation tools and statistical methods to illustrate dataset characteristics like size, quantity, and accuracy to realise a much better understanding of the character of the data. Data exploration methodologies offer both manual and automatic data exploratory research software solutions that visually explore and identify relationships between various data variables, the structure of the dataset, the existence of outliers and the distribution of understanding values can be evaluated by analysing patterns and points of interest, enabling data analysts to gain a wider knowledge. Data is typically collected in massive, unstructured volumes from multiple sources, and data analysts must first comprehend and construct a comprehensive analysis of the study before attempting to extract information.

Data Exploration in machine learning.

A Machine Learning project is only as good as the data on which it is built. Machine learning data exploration models should consume large amounts of data in order to perform well, and model accuracy suffers whether that data isn't really extensively explored first. The following are the data exploration steps that can be taken prior to actually constructing a machine learning model: Identifying variables: describe every variable as well as its involvement in the dataset. Univariate analysis: for continuous data, create box plots or histograms for each variable separately; for categorical variables, create bar charts to show frequency distributions. Bivariable analysis entails determining the interaction of variables through the development of visualisations. Scatter plots, both continuous and discontinuous stacked column chart and categorical and categorical: categorical and categorical: categorical and categorical: categorical and categorical: categoric Boxplots coupled with swarm plots for categorical and continuous data. The final aim of data exploration in machine learning will give informative or useful data and insights that will motivate successive feature engineering and building model. By developing features from raw data, feature engineering accelerates the machine learning experience and enhances the predictive machine learning algorithms.

Data pre- processing

Pre-processing of data. Simply defined, data pre-processing is a data mining approach that entails converting raw data into a format that can be understood. Real-world data is frequently inadequate, unreliable, and/or lacking in specific behaviours or patterns, as well as including numerous inaccuracies. Pre-processing data is a tried-and-true way of resolving certain problems.

Data pre- processing method include these following steps: -

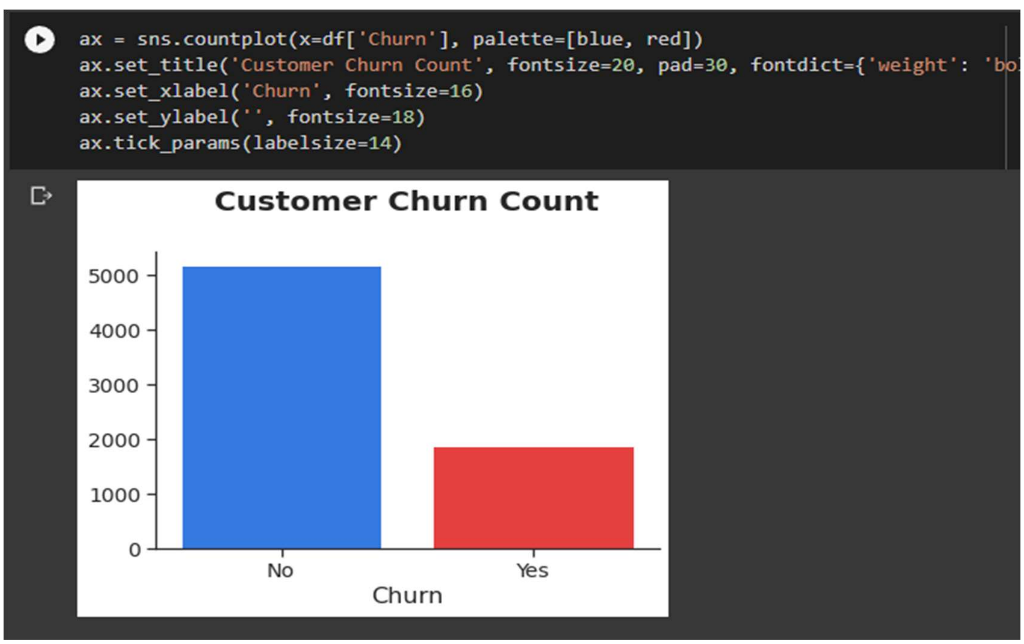
Data cleaning **Data integration** **Data reduction**

Data transformation **Data discretisation**

- **Data cleaning: -**
  Missing values are filled properly, raw data is smoothed, outliers are identified and removed, then discrepancies are resolved. Let's have a look at how I cleaned the data in our dataset. When we arranged our values in our dataset, we saw that the total charges column had missing values, therefore we'll have to clean the data using pandas. We also found that the dataset has the incorrect datatype for the columns numerical data, so we need to convert the datatype of the column. We see there are 17 category features and 4 numerical features in the dataset, yet there are just 3 numerical features in the overview. The “Total Charges “column datatype is object, but it should be floating type so firstly we convert the datatype of the Total Charges in Float.

After cleaning the data, we will set the Creating Dependent and Independent Variables from Datasets. In this dataset we had selected the independent variables as churn column, and we will now analyse each other dependent feature and see is their any relationship between two features for better insight of the data. By using Univariate Analysis and Bivariate analysis.

Imbalanced data



:- Before cleaning

We conducted analyses using Univariate Analysis, which entails analysing a single variable and examining full information true charts. The boxplot, histogram, and pie chart are the most commonly used charts for univariate analysis. We used a pie chart, a boxplot, and a histogram in this assignment. With the use of a bar plot, we performed Bivariate Analysis with dependent and independent features. After conducting our analysis, we discovered that the data in the churn column is unbalanced, which necessitates rebalancing the data. We used minmaxscaler to balance the data. We also oversample the dataset to balance it out and obtain the ideal preference. In oversampling to create a balanced dataset, oversampling is described by adding additional samples of the minority class. Whenever you will not have a lot of data to working with, oversampling could be a smart option. Whenever data analysts may not have enough data, it's a good idea to use it. One seems to be common, or the plurality, another is uncommon, or the minority. Whenever the data is limited, this strategy tries to increase the volume of rare samples to make a balance. So, this definition of less data matches for our dataset so oversampling was the perfect match for our dataset.

Exploratory data analyses

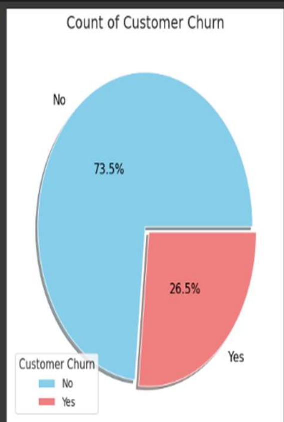
In EDA we had did univariant and bivariant analyses to get the better understanding of the dataset insights and to get the proper solution for the firm.

1. Count of churn

```
target = data['Churn'].value_counts()/len(data)*100

plt.figure(figsize = (8, 8))
plt.pie(x = target.values, labels = target.keys(), explode = [0, 0.05], autopct='%1.1f%%', shadow = True, textprops = dict(color = 'black', size = 16), colors =  ['skyblue','lightcoral'], pctdistance = 0.5)
plt.legend(title = 'Customer Churn', loc = 'lower left')
plt.title('Count of Customer Churn', fontsize = 20)
plt.show()

print(target)
```

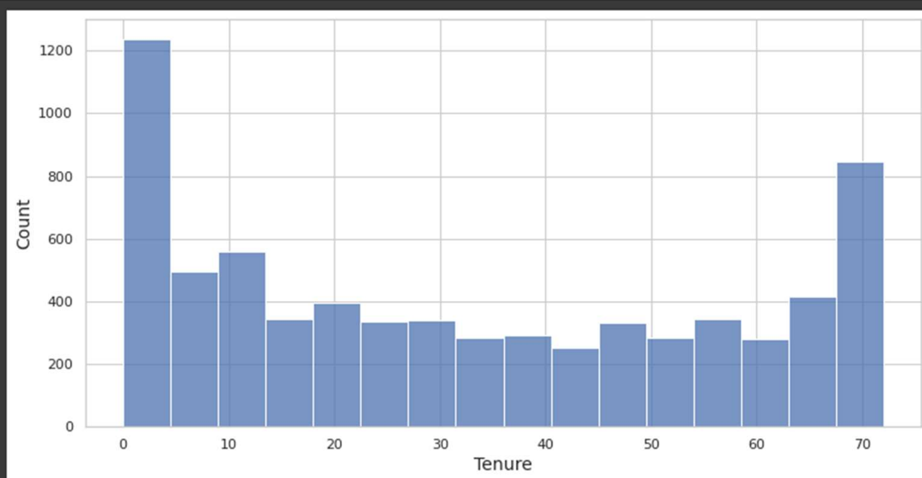



```
No     73.463013
Yes    26.536987
Name: Churn, dtype: float64
```

This above fig show the graph of the churn column and we can easily see that the data is imbalanced 73.5% is yes churn and 26.3% is no churn so it clearly an imbalanced data and we had already discussed in the above section this type of analysis is called as the univariant analyses.

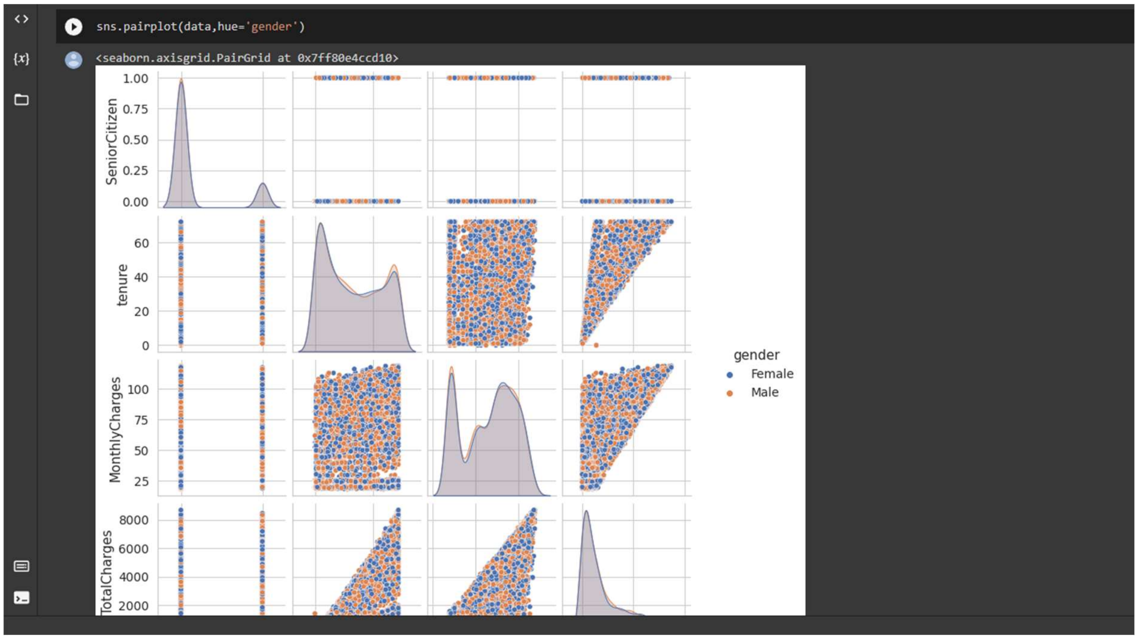
2. Count of Tenure

```
plt.figure(figsize = (12, 6))
sns.histplot(x = data['tenure'])
plt.xticks(fontsize = 11)
plt.yticks(fontsize = 11)
plt.xlabel('Tenure', fontsize = 14)
plt.ylabel('Count', fontsize = 14)
plt.show()
```



Above graph show the count of tenure and this graph is also an univariant graph and this time we plotted in histogram for a better understanding of the tenure count. And we can clearly, we see that monthly based customers are more. So, the churn customer will be also more in the monthly based tenure. This suggest that company should look into the monthly based customer and attract them into yearly package.

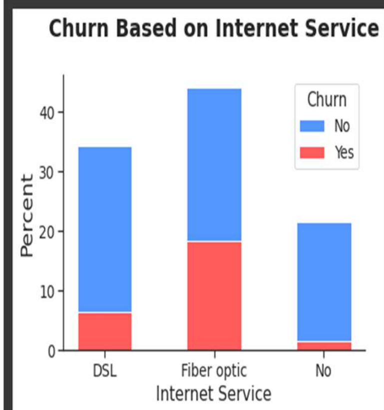
3. Pair plotting of gender



The above figure show is used to show the relation between the two variables in which we can see when gender is more churn and which gender as better loyalty

4. Churn based on Internet service.

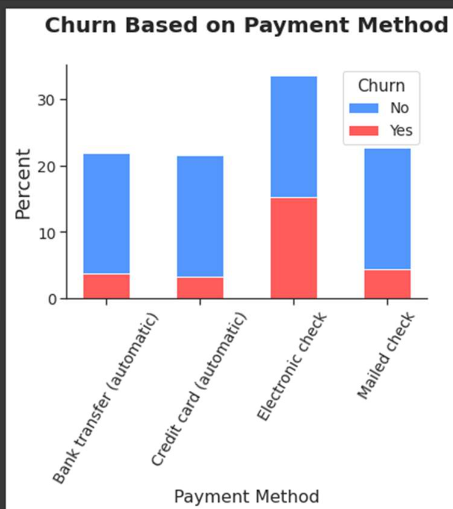
```python
ax = sns.histplot(data=df, x='InternetService', hue='Churn', shrink=0.5, multiple='stack', palette=[blue, red], stat='percent')
ax.set_title('Churn Based on Internet Service', fontsize=20, pad=30, fontdict={'weight': 'bold'})
ax.set_xlabel('Internet Service', fontsize=16)
ax.set_ylabel('Percent', fontsize=18)
ax.tick_params(labelsize=14)
```



One intriguing fact regarding payment methods is that, according to the data, all four payment methods are divided evenly among clients by 25%. Electronic checks, on the other hand, saw the most churning. There is a clear link among churning and Payment is made by electronic check. Despite the fact that the company takes electronic cheques, 57 percent of customers choose to pay with cash. Electronic cheque users have churned.

5. Churn based on Payment method

```
ax = sns.histplot(data=df, x='PaymentMethod', hue='Churn', shrink=0.5, multiple='stack', palette=[blue, red], stat='percent')
ax.set_title('Churn Based on Payment Method', fontsize=20, pad=30, fontdict={'weight': 'bold'})
ax.set_xlabel('Payment Method', fontsize=16)
ax.set_ylabel('Percent', fontsize=18)
# ax.set(ylim=(0, 2500))
ax.tick_params(labelsize=14, axis='x', rotation=60)
```



There are 1526 consumers without internet, while Fiber Optic is the most popular option, with 3096 customers. The surprising revelation found from this particular data is that nearly 45 percent of consumers utilise fibre optic, while fibre optic users have the greatest churning rate of all internet options, at about 42 percent. Even though fibre is 150 times faster than DSL, it has the greatest churn percentage. The added cost of having this service could be the explanation.

Experiment

Overview of Regression Methodology (statics)

Regression is a topic that deals with the continuous input-output system. The goal of regression is to learn a system. Learning can be measured by being able to predict the outcome of a system in relation to any input. Training is the process of creating and tuning a predictor algorithm. During training, users are given a set of known input-output pair pairs (sometimes called the data1 or simply 'the data'). This meta-algorithm is one step above the actual predictor algorithm. The output of the training algorithm is the predictor, which generates values that are like the output of the system. These training algorithms are often used to solve optimization problems. Machine-learning researchers pride their selves on creating smart learning algorithms. People who apply regression to their apps use existing training algorithms to create predictors from their data. These predictors are then used for data analysis, forecasting, optimization, and other purposes. It is often as simple as a multiplication matrix to use a predictor. This document briefly explains both how to make and use predictors to make predictions (Zohar, et al., 2011).

Why Regression?

Linear regression is required to predict churn of customer because it is difficult to analyse high-dimensional data. Linear regression uses data to create a model that can predict future data. The target is the predicted data, while the features are the predictions. Imagine, for example, trying to predict the GPA of students based on how much they studied and how classes they attended. We are interested to predict the churn of customer. y represents churn. Equation 1's x values are all "features", such as Monthly payment and tenure. Regressors are described by the theta parameters. These parameters describe how the corresponding features affect the prediction.

Overview of KNN

The supervised machine learning algorithm K-nearest neighbours (kNN) could be used to address both classification and regression problems. kNN, in my opinion, is indeed a real-world algorithm. Individuals are affected by those who surround them. Our behavior can be influenced either by people with whom we did grow up. In some respects, our families mould our personalities as well. If you really are surrounded by people that enjoy sports, it is quite possible that you might enjoy sports as well. After sure, there are exceptions. In a similar manner, kNN works. If you do have one good friend with and who you spent the majority of your time, you may find that both share common tastes and like comparable activities. This is kNN, where k=1. If you always hang out from a group of five people, every one of them seem to have an impact on your conduct, and that you will eventually become the average of the five. With k=5, this is kNN. The majority voting principle is used by the kNN classifier to identify the class of a data item. When the value of k is set to 5, the classes of the five nearest points are examined. The majority class is used to make predictions. Similarly, the mean value of the five nearest points is used in kNN regression. We watch those that are close to us, but how are data points determined the majority voting principle is employed to classify data points. When the worth of k is about to five, the classes of the five nearest points are examined. the bulk class is employed to form predictions. Similarly, the mean of the five nearest points is employed in kNN regression. We notice those who are in close proximity, but how are data points considered to be in close proximity? it's calculated the gap between data points. the gap is often measured during a kind of ways. one among the foremost widely used distance measurements is that the Euclidean distance (minkowski distance with p=2). The diagram below illustrates the way to determine the euclidean distance in an exceedingly two-dimensional space. The square of the difference between both the x and y coordinates of the locations is employed to calculate it.

Why KKN?

For the payment method predication KKN definition and implantation suits perfect as the payment as three modes and three different variables so it will group the most the used payment mode and by using this company can tackle the most popular payment mode and bring out some attractive deals to make the customer be loyal towards the company for a longer time.

Overview of K mean clustering: -

This technique is an iterative method that divides the data into K non-overlapping independent clusters or subgroups based on characteristics. It aims to maintain the clusters as widely apart as feasible while making the datasets of inter clusters as comparable as possible. If the total of the square distances between both the cluster's centroid and the data points is less than a certain threshold, the data points are assigned to a cluster. The cluster's centroid is the arithmetic mean of the data points in the cluster. Less variance in the cluster results in data points that are comparable or homogenous inside the cluster. The following requirements for the K- Means Clustering Algorithm: The number of subgroups or clusters is denoted by the letter K. x1, x2, x3, ...an = Sample or Training Set Let us just pretend we have had an unlabelled data collection which we need to partition into clusters. In this procedure, the data is separated into categories depending on several indicators, and afterwards the performance of each category is assessed. The layout of shirts in the men's apparel section at a mall, for example, is based on size criterion. It is possible to do this on the grounds of both price and brand. An optimum cluster, i.e., the value of K, would be picked.

Why used K Mean clustering.

As we need to provide a solution to the company on how to reduce churn and which customer group we should focus on more now, we created a new column where we converted the tenure of months into yearly based and yearly payment column so we can plot charge per year wise the tenure and easily identify the group of people who give us more business than those who do not. We also scatterplot it with payment type label data. So we can figure out which monthly payment set of folks we should approach and provide an appealing offer to them so they can switch from monthly to yearly.

Result and Evaluation

Machine learning classifiers have been used to determine the best model for each dataset in this paper. Each model's performance was calculated and then compared using for first problem statement where we predict the churn of the customer logistic regression with the better accuracy of 91.5%. And for the 2nd problem statement where we predicted the payment type KNN was the best fit model with the 89.9% accuracy. And to see which group customer should get the better offer to stop the churning and which group of people should be look for yearly subscription. K mean clustering was the best way to identify the group of people and see analyse it.

Conclusion

We conclude that company we offer more offer to the monthly payment customers for yearly payment which we will be benefit for the company In longer term To encourage more long-term clients, the total bill payments for a two-year contract should be reduced. The month-to-month median of each contract The month, a year, and two years are all quite near. $66, $65, and $60 are the prices. Customers made the decision to be more active. With an extra $5, you can be more flexible. Long-term contracts could be cheaper if the price is reduced. Customers who are less likely to churn, according to study. Another suggestion would be to lower the cost of fibre optic internet. Because the cost is around $70, it's possible that the high rate of churn is due to the cost. Providing a higher-quality service for a small fee might keep clients in the company longer, perhaps lowering the turnover rate.

References

Ahmad, A.K. & Jafar, A. & Aljoumaa, K. 2019. *Customer churn prediction in telecom using machine learning in big data platform. J Big Data* doi:10.1186/s40537-019-0191-6

• Campbell, Patrick. 2019. *How Involuntary Churn is Killing Business [3 Ways to Reduce in 2019]*. Available at: https://www.profitwell.com/blog/involuntary-delinquent-churnfailed-payments-recovery

• Gursoy, Simsek & Tugba Umman. 2010. *Customer churn analysis in telecomunication sector*. Journal of the School of Business Administration

• Jain, Pallav & Surana, Kushan. 2017. *Reducing churn in telecom through advanced analytics.* Available at: https://www.mckinsey.com/industries/technology-media-andtelecommunications/our-insights/reducing-churn-in-telecom-through-advancedanalytics#

• Kanniappan, Suresh. 2014. *How to Reduce Churn In A Telecom Industry.* Available at: https://www.happiestminds.com/blogs/how-to-reduce-churn-in-a-telecom/

• Staples, Kim. 2017. *Cable vs fibre optic broadband: What's the difference?.* Available at: https://www.broadbandchoices.co.uk/ask-our-expert/cable-vs-fibre

• Valchanov, Iliya. 2018. *False Positive and False Negative.* Available at: https://towardsdatascience.com/false-positive-and-false-negative-b29df2c60aca

V. Lazarov and M. Capota, ''Churn prediction,'' Bus. Anal. Course, TUM Comput. Sci, Technische Univ. München, Tech. Rep., 2007. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download? doi=10.1.1.462.7201&rep=rep1&type=pdf.

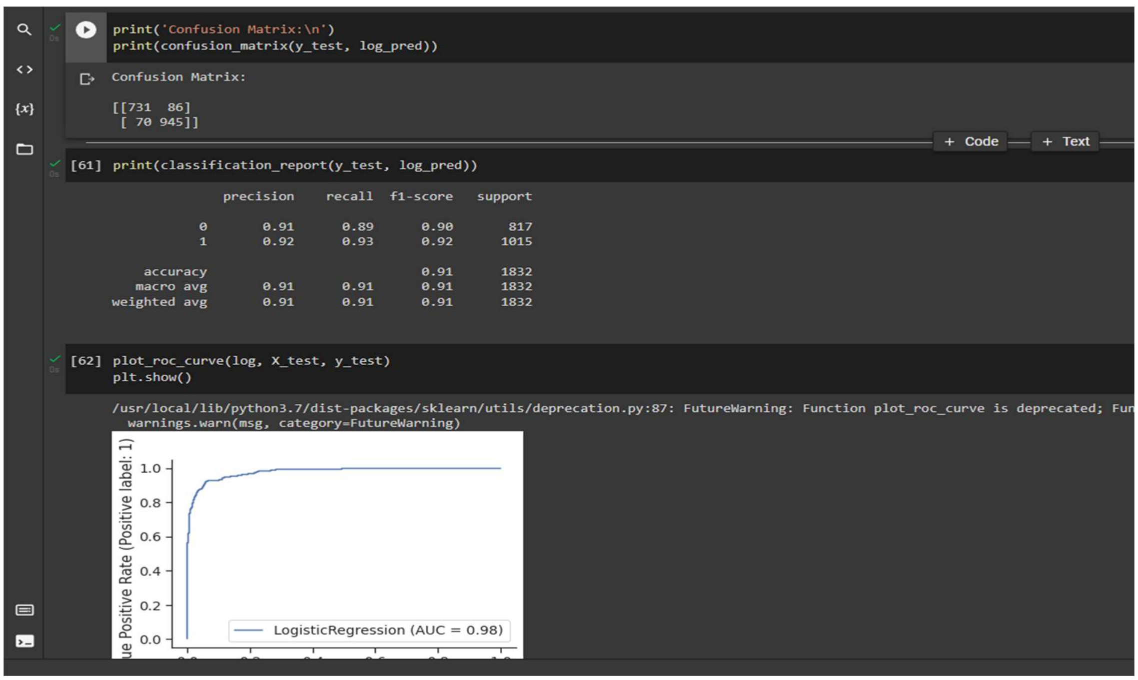
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.7201&rep=rep1&type=pdf

K- Means Clustering Algorithm | How It Works | Analysis & Implementation (educba.com)

https://www.ijert.org/research/churn-prediction-of-customer-in-telecom-industry-using-machine-learning-algorithms-IJERTV9IS050022.pdf

Appendixes:

[49] # Create a Logistic Regression object
log = LogisticRegression()
# Train the model
log.fit(X_train, y_train)
# Predict on test data
log_pred = log.predict(X_test)
print('Test Accuracy: {} %'.format(log.score(X_test, y_test).round(3)*100))
Test Accuracy: 91.5 %
logistic regression
print('Confusion Matrix:\n')
print(confusion_matrix(y_test, log_pred))
Confusion Matrix:
[[731 86]
 [70 945]]



print('Confusion Matrix:\n')
print(confusion_matrix(y_test, log_pred))
Confusion Matrix:
[[731 86]
 [70 945]]
+ Code
+ Text
[61] print(classification_report(y_test, log_pred))
precision recall f1-score support
0 0.91 0.89 0.90 817
1 0.92 0.93 0.92 1015
accuracy 0.91 1832
macro avg 0.91 0.91 0.91 1832
weighted avg 0.91 0.91 0.91 1832
[62] plot_roc_curve(log, X_test, y_test)
plt.show()
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function plot_roc_curve is deprecated; Fun
warnings.warn(msg, category=FutureWarning)
ue Positive Rate (Positive label: 1)
1.0
0.8
0.6
0.4
0.2
0.0
LogisticRegression (AUC = 0.98)

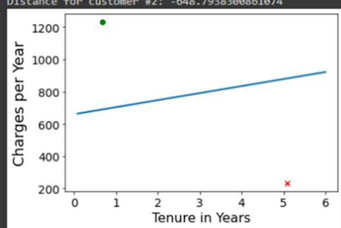
```python
ax.set_title('', fontsize=20, pad=30, fontdict={'weight': 'bold'})
ax.set_xlabel('Tenure in Years', fontsize=16)
ax.set_ylabel('Charges per Year', fontsize=18)
ax.tick_params(labelsize=14)
plt.xticks(range(0,7,1))
# plt.yticks(range(650,1000,50))

plt.scatter(x=cust1['tenureinyears'], y=cust1['yearly'], c='green', marker='o')
print('Distance for customer #1:', cust1['distance'])

plt.scatter(x=cust2['tenureinyears'], y=cust2['yearly'], c='red', marker='x')
print('Distance for customer #2:', cust2['distance'])

plt.show()
```

```
Distance for customer #1: 541.8888578283457
Distance for customer #2: -648.7938300861074
```



We plotted the two clients' data points on the graph above. We can see right away that customer one (green dot) is above the line of best fit, indicating that he or she is worth retending because the distance between the data point and the line of best fit is positive. The computed distance is the greatest discount you can offer this particular consumer. Customer two (red x-mark), on the other hand, falls below the line of best fit, indicating that retending this customer is not lucrative. The data point's distance from the line of best fit is negative.

These data points could be from consumers who have already left, but in retrospect, we could have tried to re-engage them by offering a

```python
# Create a KNN object
knn = KNeighborsClassifier()

# Train the model
knn.fit(X_train, y_train)

# Predict on test data
knn_pred = knn.predict(X_test)

print('Test Accuracy: {} %'.format(knn.score(X_test, y_test).round(3)*100))
```

```
Test Accuracy: 89.9 %
```

```python
print('Confusion Matrix:')
print(confusion_matrix(y_test, knn_pred))
```

```
Confusion Matrix:
[[314  39   6  18]
 [ 30 408  11  10]
 [  8   8 379   1]
 [ 12  16   6 363]]
```

```python
print(classification_report(y_test, knn_pred))
```

```
              precision    recall  f1-score   support

           0       0.86      0.83      0.85       377
           1       0.87      0.89      0.88       459
           2       0.94      0.96      0.95       396
           3       0.93      0.91      0.92       397

    accuracy                           0.90      1629
   macro avg       0.90      0.90      0.90      1629
weighted avg       0.90      0.90      0.90      1629
```