**Task 1**

**Problem Statement:**

The organization is experiencing a high customer churn rate across its various service offerings, including Internet, cable TV, and phone plans. The company is seeking a data-driven solution to accurately identify customers who are likely to cancel their services shortly.

**Goal:**

The goal is to develop a predictive model that can accurately predict customer churn and provide actionable insights to help the company proactively retain at-risk customers and reduce overall customer attrition.

**Data Description:**

The company has provided you with a dataset containing historical customer data, including demographic information, service usage, billing details, and customer churn status. The dataset consists of the following features:



1. **Data Gathering**

For this project, we import CSV files with pandas using the read\_csv module. The dataset is imported using the pd.read\_csv function.

1. **Data Understanding**

The custom function provides top rows, row shape, number of features, list, missing values, unique values, and data type. It shows a total missing value of 30849 based on features.

**Handling Missing Values:**

The custom function named impute\_missing\_value has been created, containing data and a message that has been updated accordingly.

Based on the business data description, imputing the missing values as follows.

1. fills in the missing values in the 'Offer' column with the string "No Offer".
2. Replace missing values in the 'Avg Monthly Long Distance Charges' column with 0 for customers with 'No' in the 'Phone Service' column.
3. Replace missing values in the 'Multiple Lines' column with "No" for customers with 'No' in the 'Phone Service' column.
4. Replace missing values in the 'Internet Type' column with the string "No Internet Type".
5. Remaining follow a similar pattern, filling in missing values in various columns like 'Avg Monthly GB Download', 'Online Security', 'Online Backup', 'Device Protection Plan', 'Premium Tech Support', 'Streaming TV', 'Streaming Movies', 'Streaming Music', 'Unlimited Data' based on the value in the 'Internet Service' column.
6. Replace missing values in the 'Cancellation Category' column with the string "No Category".
7. Replace missing values in the 'Cancellation Reason' column with the string "No Reason".

By imputing all the values, finally, there are no missing values.

**Checking the categorial data:**

The function creates a new column 'Customer Status', sets existing values to 'No Churn', creates a dictionary, maps values to numerical values, and updates the column for further analysis.

1. **Data Exploration**

It calculates unique values in the 'Customer Status' column of DataFrame df using churn\_counts(). The data is prepared for a pie chart using labels, sizes, and colors. The pie chart is created using Matplotlib, with a size of 8x8 inches and a percentage of each slice. The chart's title, axis, and display are set using plt.title, axis, and show().

The pie chart displays the distribution of 'No Churn' with 73.5% and 'Churn' with 26.5% values in the dataset, providing a quick understanding of the overall customer base churn rate.

A blue and orange pie chart

Description automatically generated

The DataFrame df uses select\_dtypes() to identify and convert categorical and numerical columns into a list, iterating through each column to check unique values.

The resulting 'Age\_bins' column will contain the bin labels for each row, based on the age value in the 'Age' column.

A white rectangular object with black lines

Description automatically generated with medium confidence

Insights: The analysis suggests that there is a gender-based difference in churn rates, with female customers being more likely to churn.

A white background with numbers

Description automatically generated

Insights: The graph suggests that customers churn when they are more likely to be unmarried.

A white rectangular object with black text

Description automatically generated

Insights: The graph indicates that zero dependents are more likely to churn rather than 1 or more dependents.

A graph with numbers and lines

Description automatically generated with medium confidence

Insights: The graph indicates that zero or one referrals are more likely to churn when compared to more referrals.

A white background with black text

Description automatically generated

Insights: This graph indicates that people with no offer are more likely to churn compared to people who are having offers.

A white background with black text

Description automatically generated

Insights: The graph indicates that people with phone services are more likely to churn than the no phone services.

A white graph with black text

Description automatically generated

Insights: The graph indicates that customers with single lines are more likely to churn when compared with multiple lines.

A white screen with black text

Description automatically generated

Insights: The graph indicates that customers with internet service are more likely to churn that the no phone service

A graph with numbers and lines

Description automatically generated with medium confidence

Insights: The graph indicates that customers with Fiber Optic Internet type are more likely to churn.

A white graph with black text

Description automatically generated

Insights: The graph indicates that customers with no online security are more likely to churn.

A white graph with black text

Description automatically generated

Insights: The graph indicates that customers with no Online Backup are more likely to churn.

A white graph with black text

Description automatically generated

Insights: The graph indicates that customers with no Device Protection are more likely to churn.

A white rectangular object with black text

Description automatically generated

Insights: The graph indicates that customers with no Premium Tech Support are more likely to churn.

A white graph with black text

Description automatically generated

Insights: The graph indicates that customers with no Streaming TV attribute are more likely to churn.

A white screen with black text

Description automatically generated

Insights: The graph indicates that customers with no Streaming Movies Attribute are more likely to churn.

A white graph with black text

Description automatically generated

Insights: The graph indicates that customers with no Streaming Music Attribute are more likely to churn.

A graph with numbers and symbols

Description automatically generated

Insights: The graph indicates that customers with no Unlimited Data Attribute are more likely to churn.

A white graph with numbers and a black text

Description automatically generated with medium confidence

Insights: The graph indicates that  month to month are more likely to Churn

A graph with numbers and symbols

Description automatically generated

Insights: The graph indicates that customers with yes Paperless Billing are more likely to churn.

A white graph with numbers

Description automatically generated

Insights: The graph shows that bank withdrawals are more likely to churn

A white graph with numbers and a black line

Description automatically generated with medium confidence

Insights: The graph provides the after binning, customers with 81-100 and 41-60 are more likely to churn compared to other ages.

A graph of a diagram

Description automatically generated with medium confidence

Insights: Two density plots are presented in the graphic; one represents the monthly charge distribution by customer churn status, while the other represents the overall charge distribution by churn status. The blue curve displays a wider range with a peak of approximately 80–100 monthly charges and a higher proportion of consumers with churn status 0 (not churned). With a peak of approximately 4000-6000 total charges, the orange curve has a wider dispersion and suggests a bigger percentage of churned clients.

**Converting Categorical to binary variable:** For further analysis, categorical columns are converted to binary.

**Outliers:**

* Identifying the Outlier columns, which need to be eliminated as a process of modeling because no imputation technique is helpful for this data.
* The following values are removed because they contain outliers: Total Refunds, Total Extra Data Charges, Total Long-Distance Charges, Total Revenue, Phone Service, Internet Service, Average Monthly GB Download, and Referrals.

**Map Coordinates (Zip Code ):**  
A map of the state of california

Description automatically generated

**Correlation matrix:**

A colorful grid with black and red squares

Description automatically generated

Based on the correlation matrix, dropping the below features:  
'Total Revenue', 'Total Charges', 'Latitude', 'Streaming Music', 'Longitude'

1. **Feature Engineering**

**Based on the above analysis dropping a few features in feature Engineering**

* Customer ID - Churn rate is not dependent on the Customer ID
* Gender - There is a small gap between male and female churn, female is more likely to chrun.
* Age - After binning, Customers with 81-100 and 41-60 are more likely to churn compare to other ages
* Married- Unmarried are more likely to churn
* Number of Dependents- Zero dependents are more likely to churn (Large difference when compared with dependents - more outliers
* City- based on the data city has a clear relationship with the Zip code, we can eliminate this
* Zip Code - based on the zip code there is a slight change in the churn and no-churn ratio
* Latitude - Since the latitude is dependent on the Zip code and City, we can eliminate
* Longitude - Since the Longitude is dependent on the Zip code and City, we can eliminate
* Number of Referrals - 0 and 1 referrals are more likely to churn when compared to more referrals - more outliers
* Tenure in Months - Tenure in months is directly related to the Contract, we can eliminate
* Offer - Customers with No offer is more likely to Churn compared to offers
* Phone Service - Customers with Phone service is more likely to churn than the no phone service - more outliers
* Avg Monthly Long-Distance Charges - Since there is a total charges data and more outliers
* Multiple Lines - Customers with a Single line are more likely to churn
* Internet Service - Customers with internet service are more likely to churn than the no phone service - more outliers
* Internet Type - Customers with Fiber Optic Internet type is more likely to churn.
* Avg Monthly GB Download-more outliers
* Online Security - Customers with No Online Security are more likely to Churn
* Online Backup - Customers with No Online Backup are more likely to Churn
* Device Protection Plan - Customers with No Device Protection Plan are more likely to Churn
* Premium Tech Support - Customers with No Premium Tech Support are more likely to Churn
* Streaming TV - Customers with No Streaming TV are more likely to Churn
* Streaming Movies - Customers with No Streaming Movies are more likely to Churn
* Streaming Music- Customers with No Streaming Music are more likely to Churn, also there exists a correlation.
* Unlimited Data - Customers with No Unlimited Data are more likely to Churn
* Contract - Month to Month are more likely to Churn
* Paperless Billing - Customers with Yes Paperless Billing are more likely to churn
* Payment Method - Bank withdrawals are more likely to churn.
* Monthly Charge - Less monthly charges are more likely to retain but 60-100 are a bit different
* Total Charges-Less total charges are more likely to be retained, according to the correlation matrix since it is positively correlated
* Total Refunds-more outliers
* Total Extra Data Charges-more outliers
* Total Long-Distance Charges--more outliers
* Total revenue outliers exist, and there exists a positive correlation
* Customer Status - Target Column.
* Cancellation Category--based on the business idea, the category has more empty values and is not useful for the model.
* Cancellation Reason based on the business idea, Reason has more empty values and is not useful for the model.

**Training:**

I have split the data into 80 % for training and 20 % for testing using the train\_test\_split () function from the sklearn.model\_selection module.

1. **Model Selection**

Logistic Regression

* Description: Statistical Model for Binary Classification Tasks, Outputting Probability Values.
* Usage: Healthcare (Disease Prediction), Marketing (Customer Churn Analysis)

K-Nearest Neighbors

* Description: Simple Algorithm Classifying Objects Based on Majority Class of Their K-Nearest Neighbors.
* Usage: Classification and Regression Tasks Where Instances Are Close in Feature Space.

Naive Bayes

* Description: Probabilistic Classifier Based on Bayes' Theorem with Independence Between Features Assumed.
* Usage: Text Classification, Spam Filtering, Recommendation Systems.

Support Vector Machine

* Description: Powerful Algorithm Finding Optimal Hyperplane to Separate Classes.
* Usage: Image Recognition, Text Classification, Bioinformatics.

Decision Tree

* Description: Recursively Partitions Data Based on Features to Make Decisions in a Tree-Like Structure.
* Usage: Predictive Modeling, Classification.

Random Forest

* Description: Ensemble Learning Method Aggregating Predictions of Multiple Decision Trees.
* Usage: Finance, Healthcare, Large Datasets, High-Dimensional Feature Spaces.

**How the hyperparameters can affect the performance of each model:**

Random Forest Classifier:

n\_estimators: The number of trees in the forest. Increasing this can improve the model's performance, but also increases the training time and the risk of overfitting.

max\_depth: The maximum depth of the trees. Limiting the depth can help prevent overfitting, but a depth that is too low may lead to underfitting.

**Logistic Regression:**

C: The inverse of the regularization strength. A smaller value of C encourages more regularization, which can help prevent overfitting.

penalty: The type of regularization to use. 'l1' (Lasso) and 'l2' (Ridge) regularization have different effects on the model's coefficients and can be used to encourage different types of sparsity.

**K-Nearest Neighbors:**

n\_neighbors: The number of neighbors to consider when making a prediction. A larger value can make the model more robust to noise but may also lead to overgeneralization.

weights: The weight function used in prediction. 'uniform' assigns equal weight to all neighbors, while 'distance' assigns more weight to closer neighbors.

**Decision Tree Classifier:**

max\_depth: The maximum depth of the decision tree. Limiting the depth can help prevent overfitting, but a depth that is too low may lead to underfitting.

min\_samples\_split: The minimum number of samples required to split an internal node. Increasing this can make the tree more robust to noise but may also lead to underfitting.

**Naive Bayes:**

Naive Bayes is a simple algorithm that does not have any hyperparameters to tune. It makes predictions based on the assumption of independence between features, which can be effective for certain types of problems.

**Below are the model scores:**

A graph showing a number of different colored squares

Description automatically generated with medium confidence

Random Forest turned out to be the best model for this problem with an accuracy score of 82% when compared with Logistic Regression, K-Nearest Neighbors, Decision Tree, and Naïve Bayes models. The following confusion matrix shows the performance of the model:

A blue squares with numbers and labels

Description automatically generated

True Positive: 855 ( No Churn ),False Positive: 130,False Negative: 245 ( Likely to Churn ),True Negative: 179

By this, we can say that 855 customers will not churn, and 245 customers will churn from the network in California.

1. **Model Validation**

For model validation, I took a random 2000 sample from the data split into the Validation data, looked at how the predicted data was working, and was able to achieve 81 % accuracy.

**The following confusion matrix shows the performance of the model:**

**A blue squares with white text

Description automatically generated**

True Positive: 277, False Positive: 49, False Negative: 51, True Negative: 23

By this, we can say that 277 customers will not churn, and 51 customers will churn from the network in California.

1. **Conclusion**

Based on the detailed analysis, here are some measures that can be taken to avoid customer churn:

**Target Specific Customer Segments:**

* Focus on retaining customers in the age groups of 41-60 and 81-100, as they are more likely to churn.
* Prioritize retaining unmarried customers and those with zero dependents, as they have a higher churn rate.

**Improve Offer and Contract Structure:**

* Offer more attractive plans and incentives to customers without any offers, as they are more likely to churn.
* Encourage customers to move away from month-to-month contracts, as they have a higher churn rate.

**Enhance Service and Support:**

* Improve the quality of phone service and internet service, as customers with these services are more likely to churn.
* Provide better online security, online backup, device protection plans, and premium tech support to retain customers.

**Optimize Content and Media Services:**

* Ensure that customers are satisfied with the streaming TV, streaming movies, and streaming music services, as the lack of these services is associated with higher churn rates.
* Offer unlimited data plans, as customers without unlimited data are more likely to churn.

**Monitor and Analyze Charges and Refunds:**

* Investigate the factors behind high total charges, extra data charges, and long-distance charges, as these are associated with higher churn rates.
* Analyze the reasons for high total refunds and understand how to minimize them.