Customer Churn Analysis for Telecommunications Company

FINAL REPORT

ACS WIL PROGRAM – GROUP 3

ROSHAN GAIRE
FAWAD IJAZ
HUZAIFA
HIMANSHU CHHETRI
TALHA SHAMIM

Contents

1. Data Preparation and Cleaning

1.1 Overview of Dataset

The dataset used for this analysis consists of a target variable, **churn**, which indicates whether a customer is likely to leave the service. Additionally, it includes several predictor variables related to customer characteristics, service usage, contract terms, and tenure with the telecommunications company. The objective of the data preprocessing stage was to ready the dataset for accurate clustering and predictive analysis, ensuring each variable is in a format compatible with machine learning algorithms.

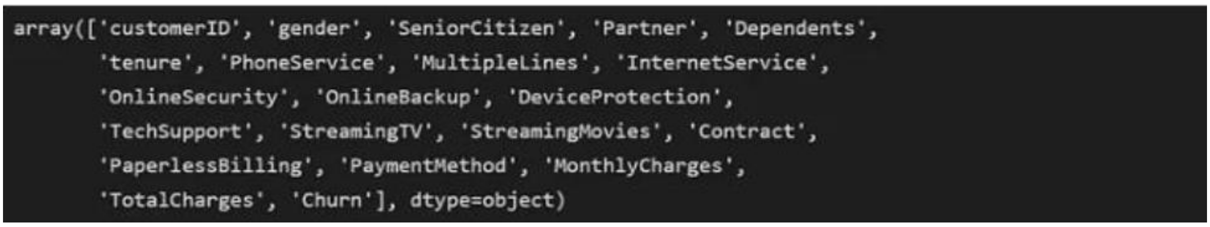
The dataset consists of **7043 rows and 21 columns**, each row representing a customer. Below are the dataset attributes:

Code:

```
import pandas as pd
# Load the dataset
df = pd.read_csv('Customer-Churn.csv')
# Display shape and column names
print(df.shape)  # Output: (7043, 21)
print(df.columns.values)
```

Output:

This displays an array of the column names:



```
array(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents',
       'tenure', 'PhoneService', 'MultipleLines', 'InternetService',
       'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',
       'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',
       'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges',
       'TotalCharges', 'Churn'], dtype=object)
```

- CustomerID: Unique value
- Gender: Whether the customer is a male or a female
- SeniorCitizen: Whether the customer is a senior citizen or not (1, 0)
- Partner: Whether the customer has a partner or not (Yes, No)
- Dependents: Whether the customer has dependents or not (Yes, No)
- tenure: Number of months the customer has stayed with the company
- PhoneServices: Whether the customer has a phone service or not (Yes,No)
- Multipleline: Whether the customer has multiple lines or not (Yes, No, No phone service)
- InternetService: Customer's internet service provider (DSL, Fiber optic, No)
- OnlineSecurity: Whether the customer has online security or not (Yes, No, No internet service)
- OnlineBackup: Whether the customer has online backup or not (Yes, No, No internet service)
- DeviceProtection:Whether the customer has device protection or not (Yes, No, No internet service)

- TechSupport: Whether the customer has tech support or not (Yes, No, No internet service)
- StremingTV: Whether the customer has streaming TV or not (Yes, No, No internet service)
- StreamingMovies: Whether the customer has streaming movies or not (Yes, No, No internet service)
- Contract: The contract term of the customer (Month-to-month, One year, Two year)
- PaperlessBilling: Whether the customer has paperless billing or not (Yes, No)
- PaymentMethod:The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
- MonthlyCharges: The amount charged to the customer monthly
- TotalCharges: The total amount charged to the customer
- Churn:Whether the customer churned or not (Yes or No)

1.2 Handling Missing Values

In real-world datasets, missing values are common and, if unaddressed, can introduce bias into the analysis. For this dataset, any missing values in numerical columns were replaced by the mean value of that column. For example, if a customer's data for "total monthly charges" was missing, it was replaced by the average monthly charges across all customers. This approach minimizes the loss of information and helps retain the integrity of the dataset by maintaining consistency across important features.

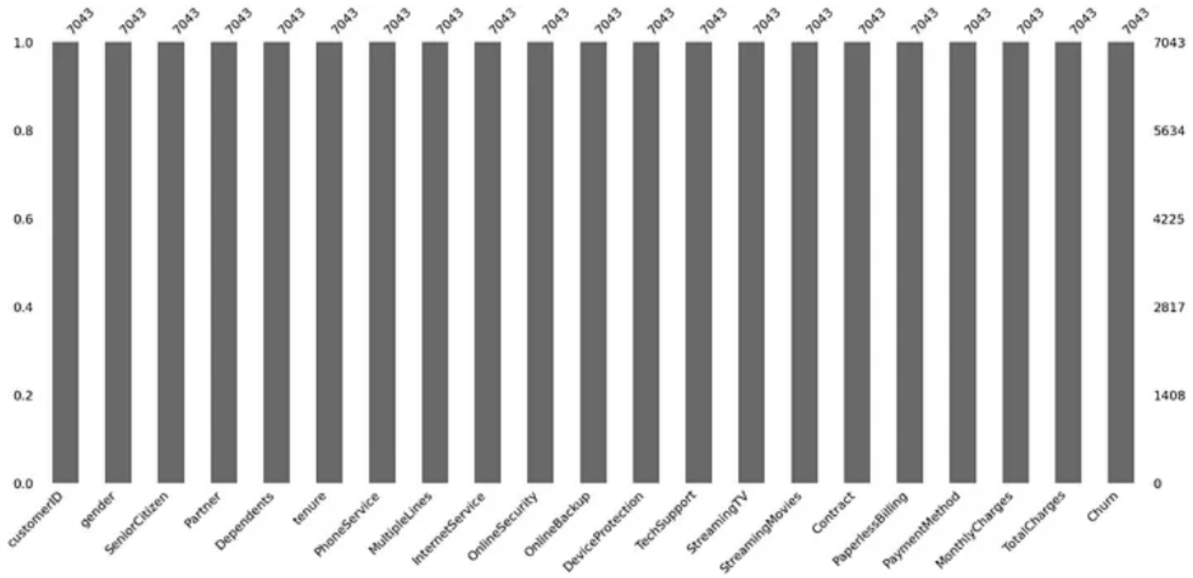
To ensure the dataset is complete and suitable for analysis, missing values were visualized and addressed:

Code:

```
import missingno as msno
import matplotlib.pyplot as plt
# Visualize missing values
msno.bar(df)
plt.show()
```

Output: **Missing Values Visualization (Bar Chart)**

- **Details**: The bar chart illustrates the presence and extent of missing data across columns in the dataset.
- **Explanation**:
  - Missing data, especially in numerical fields like TotalCharges, were identified and addressed through imputation.
  - The imputation technique used was replacing missing values with the column mean, which preserves dataset integrity and prevents data loss.
  - This preprocessing step is critical for ensuring that machine learning algorithms can train effectively without being skewed by incomplete data.



1.3 Encoding Categorical Variables

Machine learning models require numerical inputs, so categorical variables—such as contract type, payment method, and internet service—were converted into numeric representations. Label encoding was applied, which assigns an integer to each unique category. For instance, "Month-to-Month" contracts were encoded as "0", "One-Year" as "1", and "Two-Year" as "2". This transformation allows the model to interpret and process these categorical features effectively within the analysis.

1.4 Feature Scaling

Given that features vary in their scales, it was essential to apply feature scaling to ensure balanced contributions across variables. For instance, “monthly charges” range between $20 and $150, while “tenure” spans from 1 to 72 months. Using **StandardScaler**, we normalized these features to a common scale. This scaling process standardizes the dataset, enabling the model to learn efficiently without being disproportionately influenced by any single variable due to its range.

1.5 Splitting the Dataset

To evaluate the predictive model’s performance accurately, the dataset was divided into training and testing sets. We used 80% of the data for training, allowing the model to learn patterns from this subset, and reserved 20% for testing. This split ensures that the model's performance can be validated on unseen data, providing a realistic assessment of its effectiveness when applied to new data.

2. Exploratory Data Analysis and Visualization

2.1 Churn Distribution

The target variable Churn indicates whether a customer left the service. The dataset has:

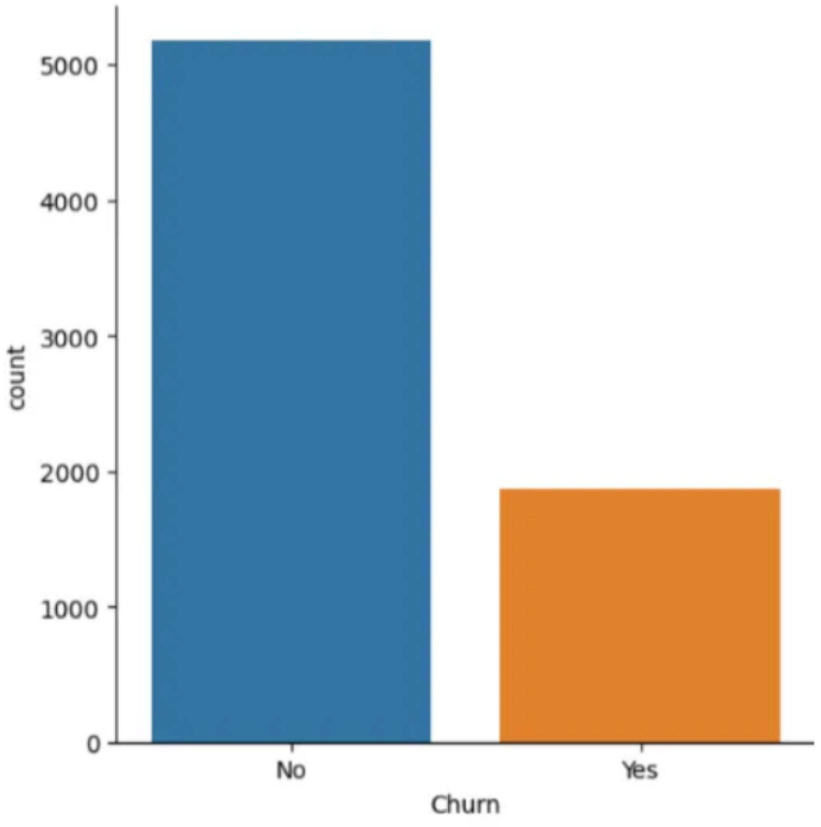
Code:

```
# Count the values for Churn
print(df['Churn'].value_counts())
sns.catplot(data=df, x="Churn", kind="count");
plt.title('Distribution of Churn')
plt.show()
```

Output: Churn Distribution (Bar Chart)

- **Details**: Displays the count of customers who churned versus those who didn’t.
- **Explanation**:
    - The chart shows a significant class imbalance, with 5174 non-churned customers and only 1869 churned customers.
    - This imbalance can negatively affect the predictive model's performance, as the model might bias towards the majority class (non-churn).
    - Techniques like oversampling the minority class (churned customers) or using synthetic data generation methods (e.g., SMOTE) may be needed to address this issue.

Interpretation: The high disparity suggests churn is imbalanced, potentially requiring rebalancing techniques like SMOTE during model training.



2.2 Gender Distribution

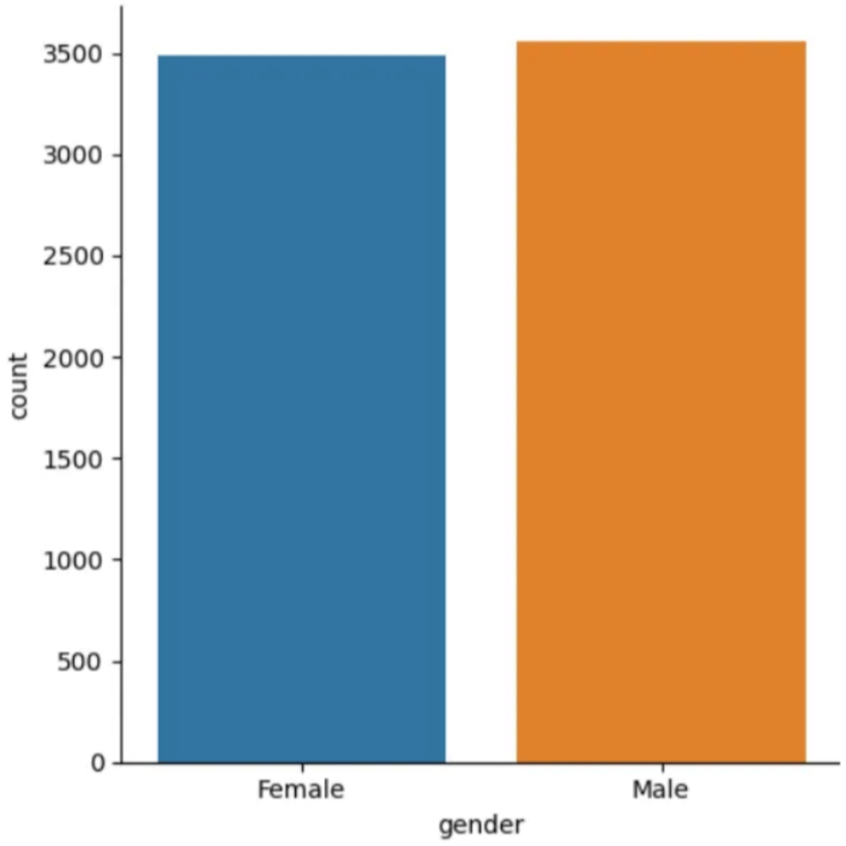
The gender distribution was visualized to assess differences in churn by gender:

Code:

```
sns.catplot(data=df, x="gender", kind="count");
plt.title('Gender Distribution')
plt.show()
```

Output: Gender Distribution (Bar Chart)

- **Details**: Compares the distribution of male and female customers in the dataset.
- **Explanation**:
  - The chart indicates a nearly equal distribution of male and female customers, suggesting no significant gender bias in the dataset.
  - This distribution helps ensure that gender-related patterns can be fairly analyzed without overrepresentation of one group.



2.3 Senior Citizen Analysis

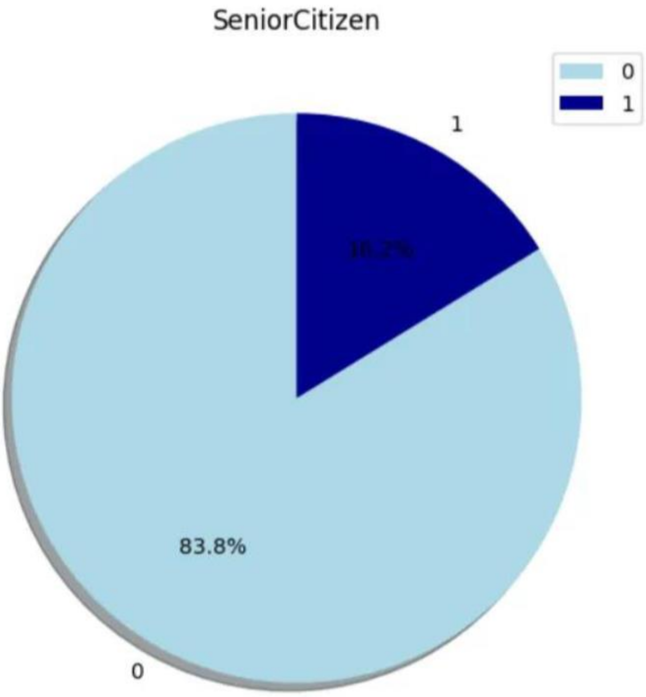
Senior citizens formed a smaller proportion of customers but had higher churn rates:

Code:

```
# Pie chart for Senior Citizen distribution
df['SeniorCitizen'].value_counts().plot(
  kind='pie', autopct='%1.1f%%', shadow=True, startangle=90)
plt.title('Senior Citizen Distribution')
plt.show()

# Histogram for churn by Senior Citizen
import plotly.express as px

fig = px.histogram(df, x="Churn", color="SeniorCitizen",
          barmode="group", color_discrete_map={"Yes": 'blue', "No": 'lightblue'})
fig.update_layout(title='Churn Distribution by Senior Citizen')
fig.show()
```

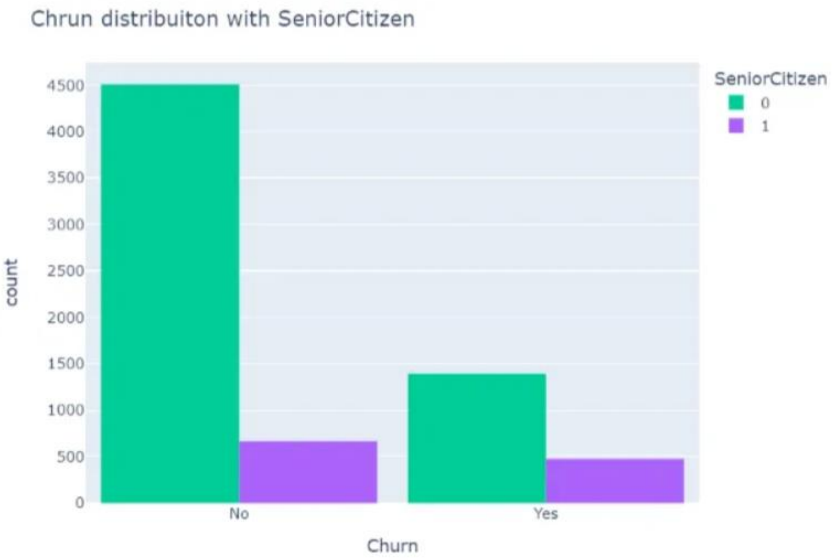
Output 1: **Senior Citizen Analysis (Pie Chart and Histogram)**

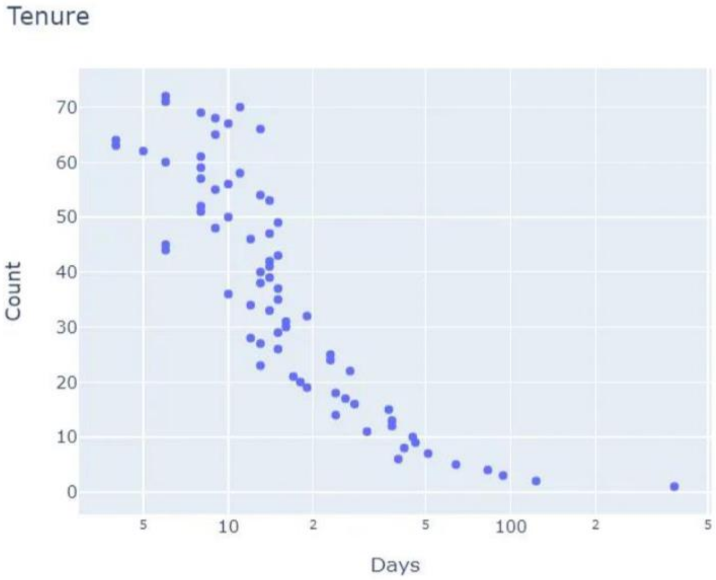
- **Details**:
  - Pie Chart: Senior citizens constitute 16.2% of the customer base.
  - Histogram: Senior citizens show higher churn rates than non-senior customers.
- **Explanation**:
  - The higher churn rate among senior citizens could point to dissatisfaction with specific services or difficulty navigating technology.
  - Targeted strategies like simplified plans or customer support specifically for senior citizens could help mitigate churn in this group.



Output 2: Tenure Distribution (Histogram and Scatter Plot)

- **Details**:
  - Histogram: Tenure distribution shows most churned customers have tenures of less than 12 months.
  - Scatter Plot: Focuses on the tenure range of churned customers.
- **Explanation**:
  - Customers with shorter tenure (<12 months) are more likely to churn, suggesting that early-stage engagement is crucial.
  - Personalized onboarding programs and early feedback mechanisms can enhance customer satisfaction during the initial months and reduce churn risks.





2.4 Internet Service and Churn

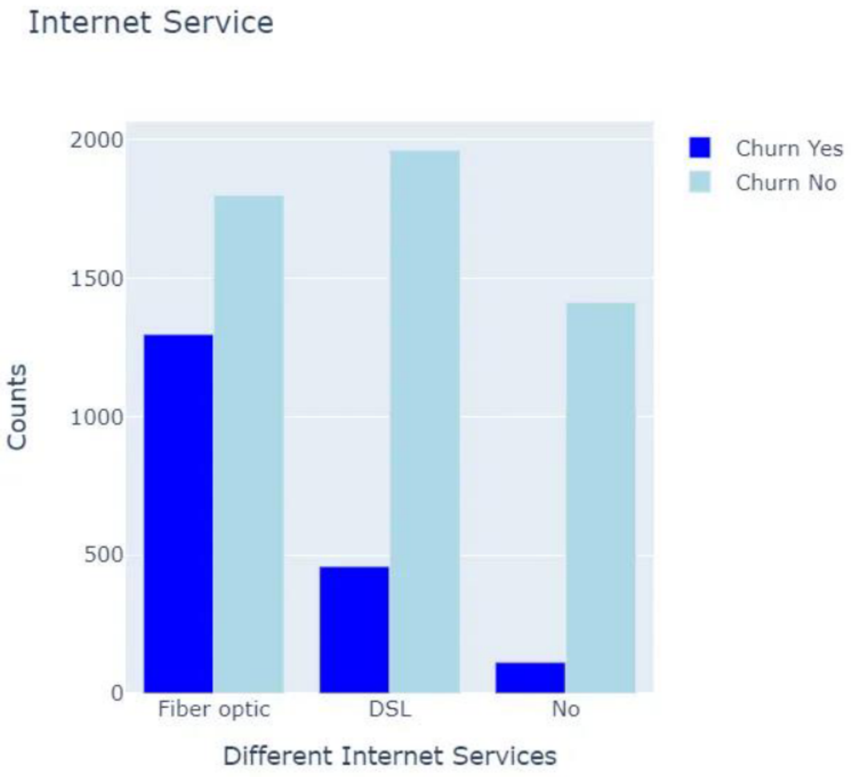
Fiber optic customers exhibited higher churn rates compared to DSL users:

Code:

```
import plotly.graph_objects as go
# Create bar plot for Internet Service distribution
fig = go.Figure()
fig.add_trace(go.Bar(name='Churn Yes', x=['Fiber optic', 'DSL', 'No'],
              y=[1297, 459, 113], marker_color='blue'))
fig.add_trace(go.Bar(name='Churn No', x=['Fiber optic', 'DSL', 'No'],
              y=[1799, 1962, 1413], marker_color='lightblue'))
fig.update_layout(title='Internet Service and Churn',
            xaxis_title='Internet Service Type', yaxis_title='Count')
fig.show()
```

Output: **Internet Service and Churn (Bar Chart)**

- **Details**: Shows churn rates for different internet service types (Fiber optic, DSL, No service).
- **Explanation**:
  - Customers using Fiber optic services show the highest churn rate, possibly due to dissatisfaction with pricing or service quality.
  - DSL users have lower churn rates, indicating better perceived value or fewer service issues.
  - Addressing Fiber optic customers' concerns through surveys or service improvements could enhance retention.



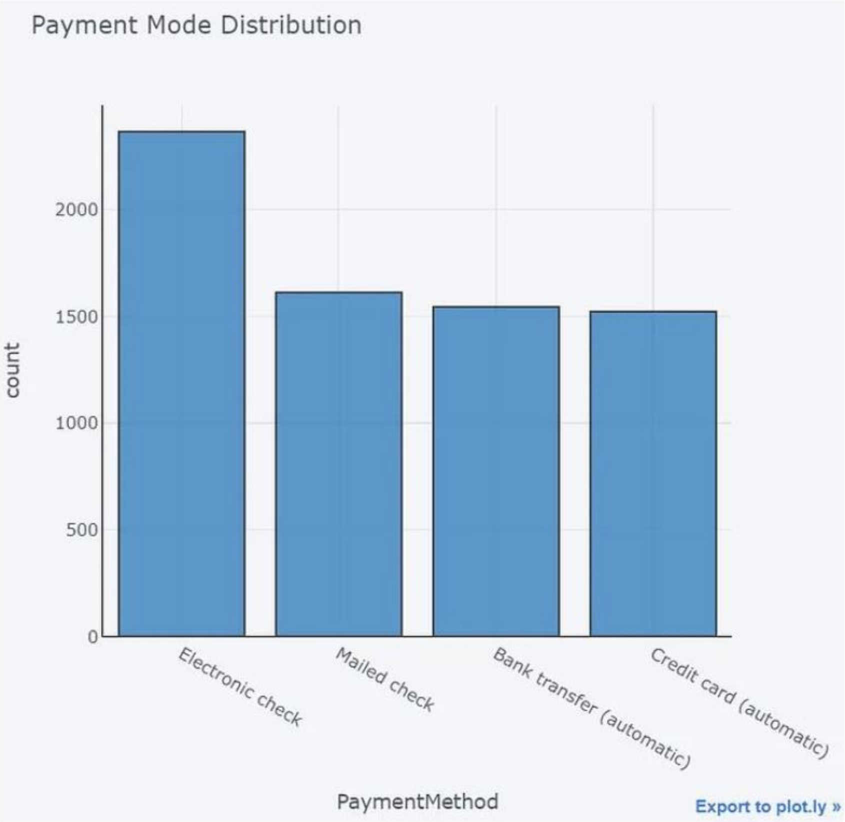
2.5 Payment Mode Distribution

The distribution of payment methods reveals insights into customer preferences and potential churn trends.

Code:

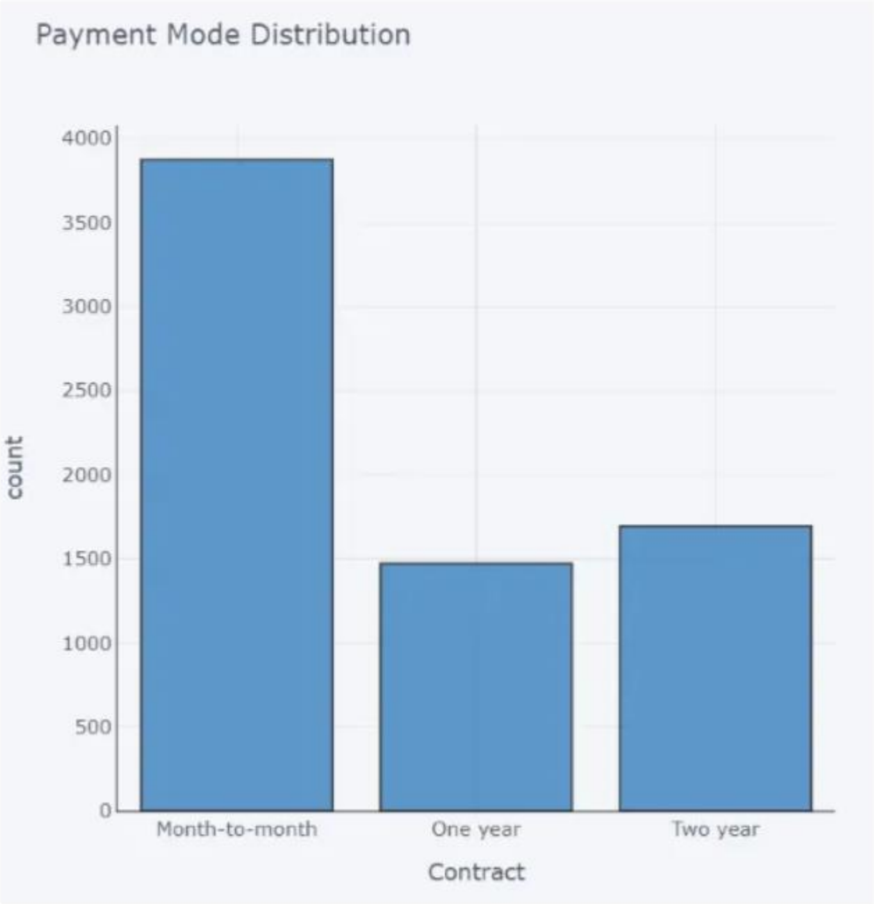
```
cf.go_offline()
cf.set_config_file(offline = False, world_readable = True)df['PaymentMethod'].iplot(kind='hist',
xTitle='PaymentMethod',
linecolor='black',
yTitle='count',
dimensions = (600,600),
title='Payment Mode Distribution')
```

Output: Running the above code gives us the following result:



Code:

```
df['Contract'].iplot(kind='hist',
xTitle='Contract',
linecolor='black',
yTitle='count',
color='blue',
size=20,
dimensions=(600,600),
title='Payment Mode Distribution')
```



2.6 Contract Distribution

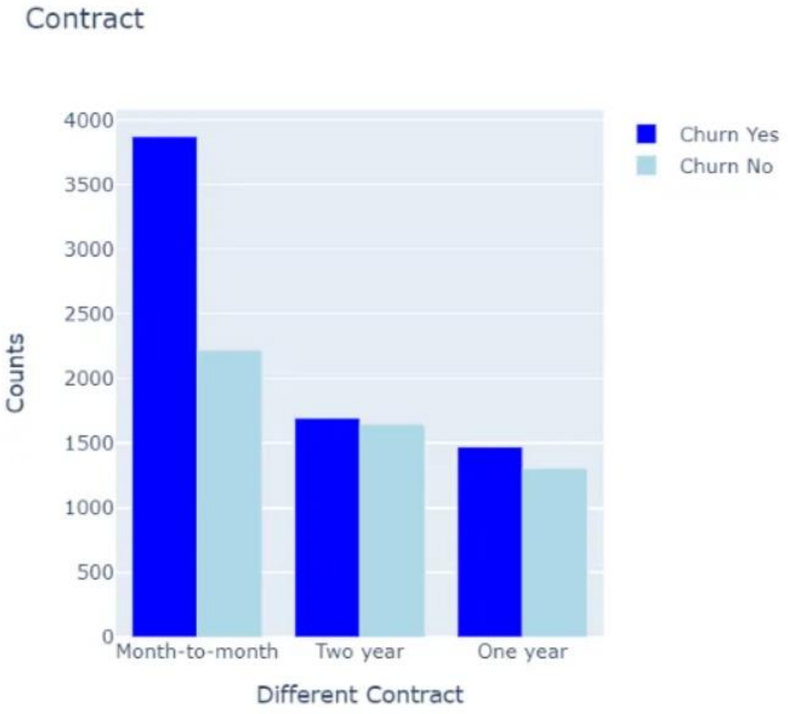
The type of contract significantly impacts customer retention.

Code:

```
fig= go.Figure()#Churn_yes
fig.add_trace(go.Bar(name='Churn Yes',
x=['Month-to-month', 'Two year', 'One year'],
y=[3875, 1695, 1473],
marker_color='blue'))#Churn_no
fig.add_trace(go.Bar(name='Churn No',
x=['Month-to-month', 'Two year', 'One year'],
y=[2220, 1647, 1307],
marker_color='lightblue'))fig.update_layout(title='Contract',
autosize=False,
width=500,
height=500)
fig.update_xaxes(title='Different Contract')
fig.update_yaxes(title='Counts')
fig.show()
```

Output: Contract Distribution (Bar Chart)

- **Details**: Compares churn rates across contract types (Month-to-month, One-year, Two-year).
- **Explanation**:
    - Month-to-month contracts exhibit the highest churn due to their flexibility.
    - Long-term contracts (One-year, Two-year) have significantly lower churn rates, as these customers likely perceive greater value or commitment.
    - Offering incentives for customers to transition from month-to-month contracts to long-term contracts can effectively reduce churn.



Interpretation:
Customers with long-term contracts are less likely to churn. Encouraging month-to-month customers to switch to annual contracts with benefits may help reduce churn.

2.7 Phone Service Distribution

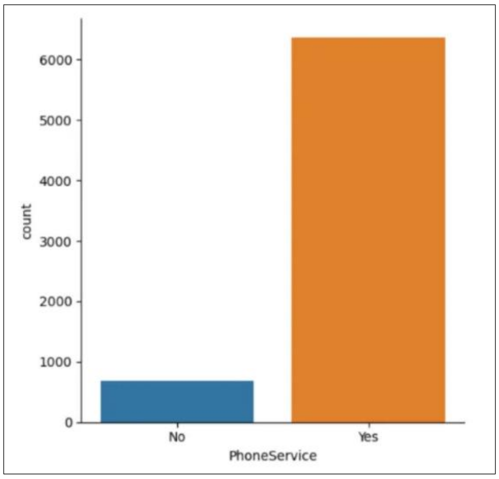
Examining Phone Service enrollment can provide insights into churn patterns.

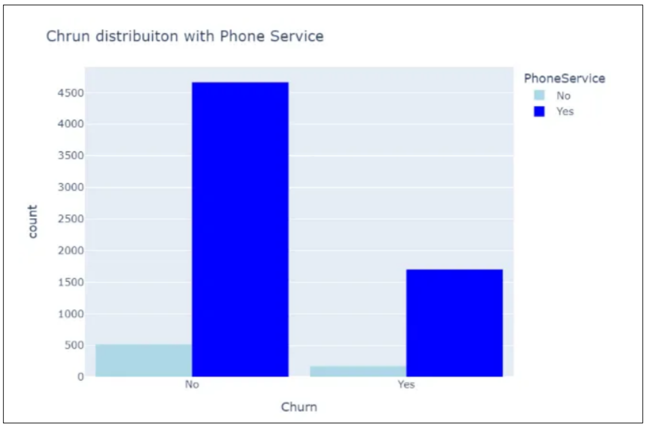
Code:

```
# Visualize Phone Service distribution
sns.catplot(data=df, x="PhoneService", kind="count");
plt.title('Phone Service Distribution')
plt.show()

# Histogram of churn distribution with Phone Service
color = {"Yes": 'blue', "No": 'lightblue'}
import plotly.express as px

fig = px.histogram(df, x="Churn",
             color="PhoneService",
             barmode="group",
             color_discrete_map=color)
fig.update_layout(title='Churn Distribution by Phone Service',
            width=700, height=500, bargap=0.1)
fig.show()
```

Output: **Phone Service and Churn (Histogram)**

- **Details**: Compares churn rates for customers with and without phone service.
- **Explanation**:
    - Customers without phone service have a smaller base but higher churn rates, suggesting they may perceive less value in the service.
    - Bundling phone services with internet or other plans at a discounted rate could address this gap and improve retention.

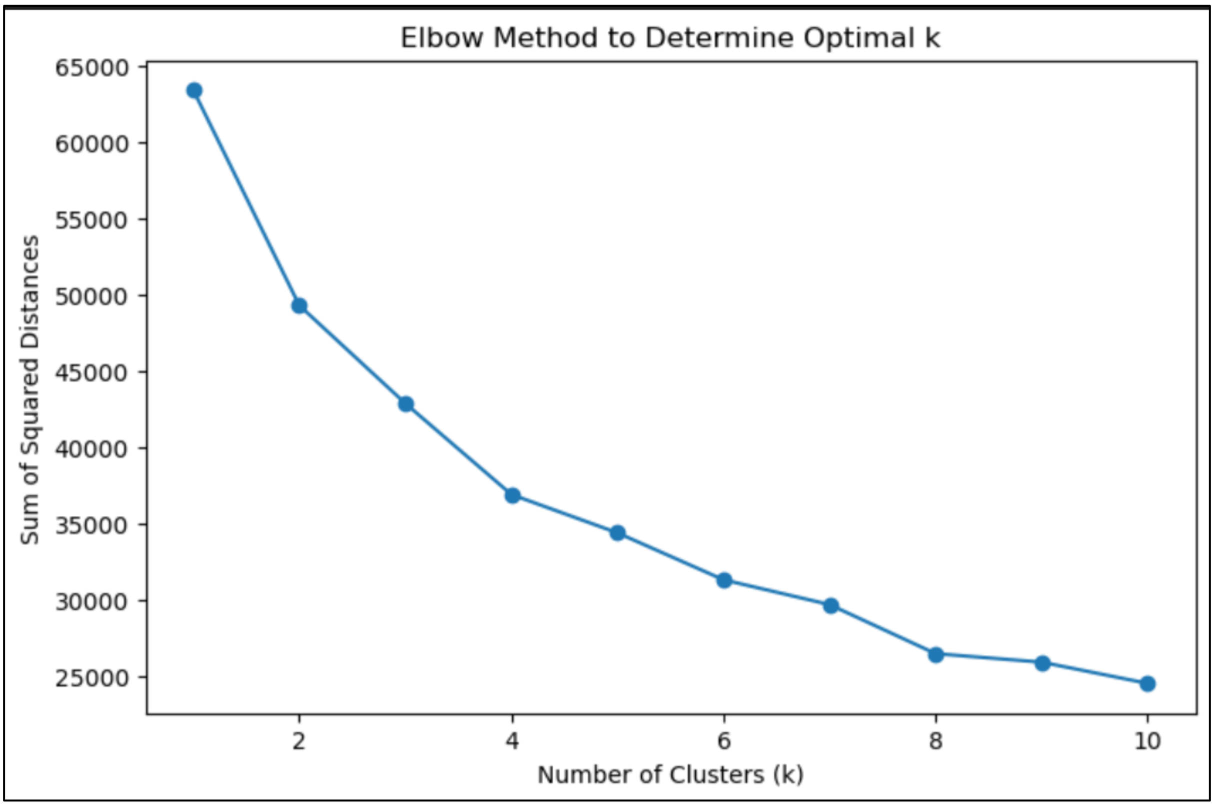




3. Determining Optimal Cluster Count

To identify the optimal number of clusters for segmenting the customer data, we employed the **Elbow Method**. This method is commonly used in clustering analysis to help determine the most effective number of clusters, balancing the level of detail captured in the segmentation without unnecessary complexity.

- **Elbow Method**: The Elbow Method involves plotting the Sum of Squared Distances (SSE) within each cluster against a range of potential cluster counts, denoted by k. As the number of clusters increases, the SSE decreases, as clusters become smaller and more closely aligned with individual data points. The point where the rate of decrease sharply slows down, forming an "elbow," typically indicates the optimal k, where adding further clusters provides diminishing returns.
- **Results**: As illustrated in the elbow graph, there is a noticeable "elbow" at k=3. Beyond three clusters, the reduction in SSE becomes more gradual, suggesting that additional clusters would add complexity without significantly improving the clustering quality. Based on this observation, we selected **three clusters** as the optimal number for segmenting our customer data. This decision allows for effective segmentation while preserving interpretability.

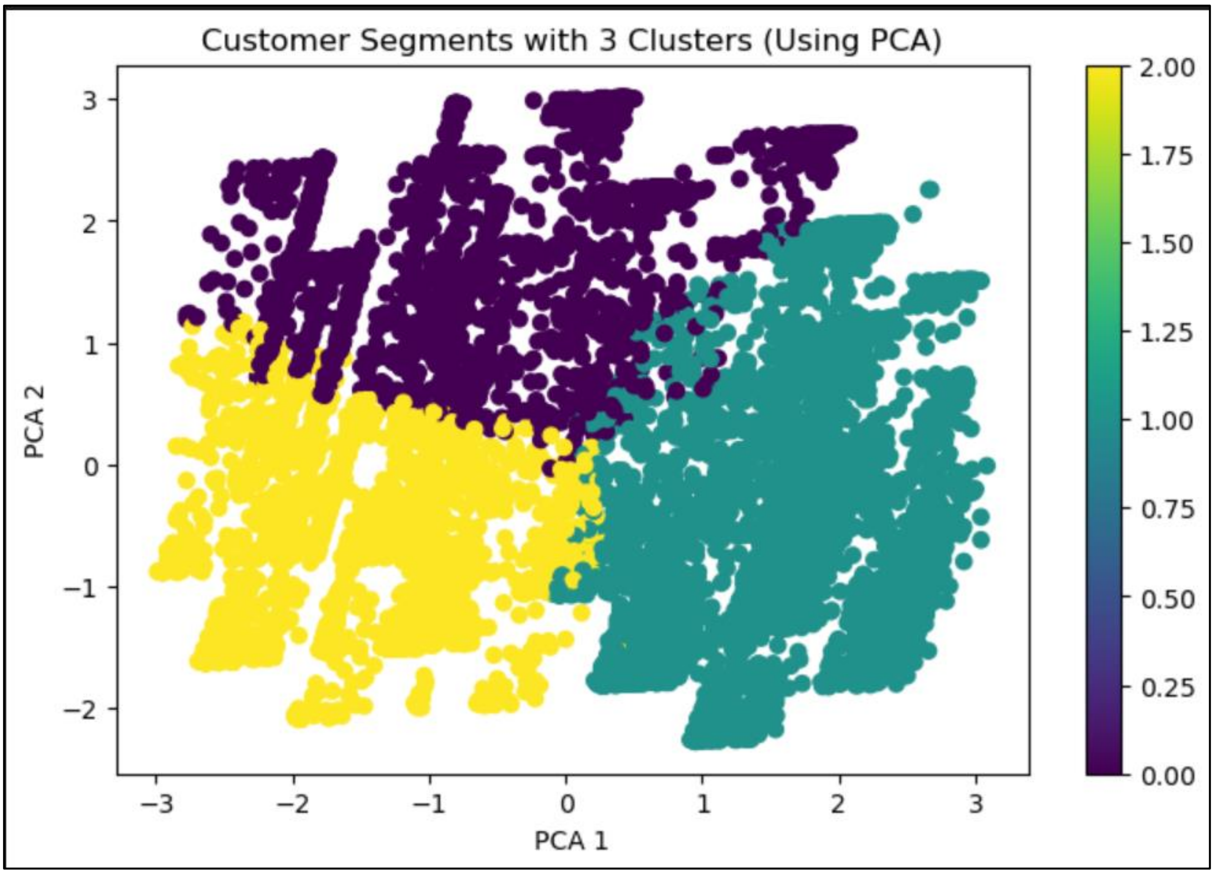


4. Cluster Analysis and Visualization

With the optimal cluster count determined, we proceeded to segment the customers into three clusters using the **KMeans Clustering Algorithm**. This approach groups customers with similar characteristics, allowing us to identify distinct segments based on their attributes.

- **Cluster Profiles and Insights**:
    - **Cluster 0** ("Low Tenure, Low Monthly Charges"): This cluster consists of customers who have shorter tenures and lower monthly charges. These customers are likely new or less engaged with the service, possibly opting for basic plans. Due to their limited engagement, they may have a higher risk of churn. Early interventions, such as personalized offers or introductory promotions, could help retain this group.
    - **Cluster 1** ("Moderate Tenure, Higher Monthly Charges"): This segment includes customers with moderate tenure and higher monthly charges, likely indicating they are regular users on premium plans. These customers might be sensitive to pricing and could respond well to loyalty programs or discounts to maintain engagement.
    - **Cluster 2** ("High Tenure, Varying Monthly Charges"): Customers in this group have been with the service for a long time and exhibit varying monthly charges, possibly due to switching plans over time. This cluster suggests loyalty, but retention strategies should focus on meeting their evolving needs through customized service options to keep them satisfied.
- **Visualization**:
    - To visualize the clusters, we used **Principal Component Analysis (PCA)** to reduce the feature dimensions to two principal components, allowing for a clear 2D visualization of the clusters. In the scatter plot, each cluster is represented by a different color, with:
        - **Cluster 0** displayed in purple,
        - **Cluster 1** in yellow, and
        - **Cluster 2** in teal.
    - This visualization demonstrates the separation between clusters, illustrating the unique characteristics of each segment and providing a clear view of customer distribution across the clusters.
- **Interpretation**: By identifying and analysing these distinct customer segments, the company can tailor retention strategies accordingly:
    - **Cluster 0** might benefit from improved onboarding experiences and special offers to enhance early engagement.
    - **Cluster 1** could respond to loyalty rewards or price-based incentives to reinforce their commitment to premium plans.

  - **Cluster 2** might appreciate personalized service adjustments or customized plans to maintain their satisfaction and long-term loyalty.



5. Feature Scaling Documentation

To ensure uniform influence across features, scaling was applied to the numerical attributes using Scikit-learn's StandardScaler.

- **Importance of Scaling**: Scalability prevents features with the large range, such as monthly charges. It disproportionately affects models compared to the smaller-range features, such as tenure.

- **StandardScaler**: This method adjusts features to have a mean of zero and a unit variance, ensuring balanced influence during training.

6. Predictive Modelling with Artificial Neural Network (ANN)

After data preparation, we designed an Artificial Neural Network (ANN) to predict customer churn. The ANN model structure and training process are as follows:

6.1 Model Architecture

The architecture of the ANN was kept relatively simple to prevent overfitting and achieve efficient learning. It includes:

- **Input Layer**: The input dimension corresponds to the number of features in the dataset after preprocessing.
- **Hidden Layers**: Two dense hidden layers were added. The first hidden layer has 16 neurons, and the second has 8 neurons, both utilizing the ReLU activation function. ReLU helps the model capture non-linear relationships in the data effectively.
- **Output Layer**: A single neuron with a sigmoid activation function was used to output the probability of churn. The sigmoid function restricts the output to a range between 0 and 1, ideal for binary classification tasks.

6.2 Model Compilation and Training

The model was compiled with:

- **Optimizer**: Adam, a commonly used optimization algorithm in deep learning that adapts learning rates during training.
- **Loss Function**: Binary cross-entropy, suitable for binary classification, helping the model minimize errors in predicting churn versus non-churn.
- **Metrics**: Accuracy was chosen as the metric to monitor the model's performance during training.

The model was trained for 50 epochs with a batch size of 32. Training was performed on the training dataset, and a portion of data was reserved for testing to evaluate the model's performance.

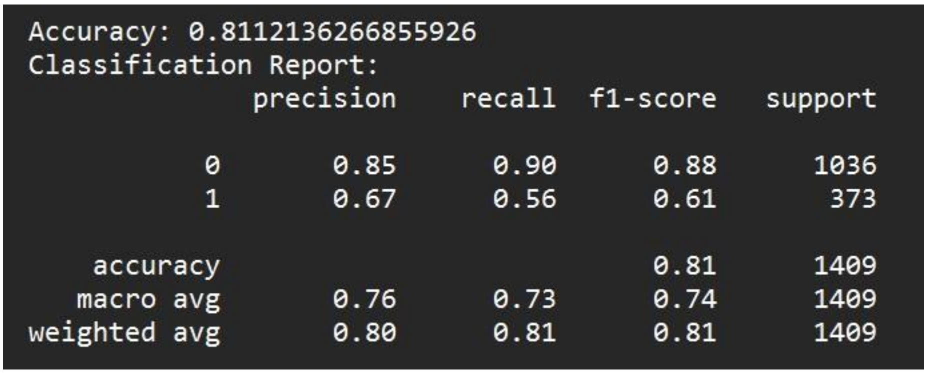
6.3 Model Evaluation

To assess the ANN’s effectiveness, we used three key evaluation metrics:

- **Accuracy**: This metric indicates the percentage of correct predictions on the test data.
- **Confusion Matrix**: Provides detailed insights into the model's classification performance by showing true positives, true negatives, false positives, and false negatives.
- **Classification Report**: Includes precision, recall, and F1-score for each class (0: No churn, 1: Churn), giving a comprehensive view of the model's ability to predict both churn and non-churn customers accurately.

After training, the model was evaluated on the test set. Below are the performance results:

Evaluation Results:



```
Accuracy: 0.8112136266855926
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.90      0.88      1036
           1       0.67      0.56      0.61       373

    accuracy                           0.81      1409
   macro avg       0.76      0.73      0.74      1409
weighted avg       0.80      0.81      0.81      1409
```

Accuracy:

- The model achieved an accuracy of approximately 80.20%, indicating that the model correctly classified around 80.20% of instances in the test data. This shows a good level of overall performance in distinguishing between customers who will churn and those who won't.

Confusion Matrix:

- The confusion matrix breaks down the number of correct and incorrect predictions for each class:
  - True Negatives (Class 0, correctly predicted no-churn customers): 893
  - False Positives (Class 0 predicted as churn): 143
  - False Negatives (Class 1 predicted as no-churn): 139
  - True Positives (Class 1, correctly predicted churn customers): 234

Classification Report:

- The classification report provides detailed metrics for both classes (No churn and Churn):
  - Precision: Measures the accuracy of the positive predictions.
    - Class 0 (No churn): 0.87
    - Class 1 (Churn): 0.62
  - Recall: Measures the ability to find all the positive samples.
    - Class 0 (No churn): 0.86
    - Class 1 (Churn): 0.63
  - F1-Score: The harmonic mean of precision and recall, providing a single metric that balances both.
    - Class 0 (No churn): 0.87
    - Class 1 (Churn): 0.63
  - Support: The number of actual occurrences of each class in the test data.
    - Class 0 (No churn): 1036
    - Class 1 (Churn): 373

- Overall Metrics:
    - Accuracy: 0.80
    - Macro Average: The unweighted mean of the precision, recall, and F1-score for all classes.
        - Precision: 0.75
        - Recall: 0.75
        - F1-Score: 0.75
    - Weighted Average: The weighted mean of precision, recall, and F1-score, considering the support of each class.
        - Precision: 0.80
        - Recall: 0.80
        - F1-Score: 0.80

Interpretation of Results

- Strengths: The model performs well in predicting non-churn customers (Class 0), with high precision and recall values. This indicates that the model is effective in correctly identifying customers who are not likely to churn.
- Areas for Improvement: The recall for churn customers (Class 1) is lower, meaning the model misses some true churn cases. While the precision and recall for churn prediction could be improved, the overall accuracy and precision make the model a valuable tool for identifying customers at risk of churn.

6.4 Interpretation with Next Steps

The ANN model effectively captures patterns in customer behavior that indicate churn risk. While the accuracy is high, the model's performance could be further improved by balancing the dataset or tuning hyperparameters to improve recall for the churn class. Additionally, incorporating methods like SMOTE (Synthetic Minority Over-sampling Technique) or adjusting class weights could help enhance recall for churn cases, thus providing a more balanced prediction model.

7. Key Drivers of Customer Churn and Retention

The analysis highlighted several influential factors in customer churn:

1. **Short Tenure (New Customers)**: New customers are more likely to churn. It emphasizes the need for early customer engagement and efforts to create satisfaction.
2. **Higher Monthly Charges**: Customers of higher priced plans are more likely to churn. This indicates the perceived value relative to cost.
3. **Service Type**: Subscribers on basic plans displayed higher churn rates, which may be due to limited-service features.
4. **Perceived Value**: Dissatisfaction with the cost-benefit ratio is a key factor in turnover. It emphasizes the need for a pricing strategy that is consistent with value.

8. Recommended Retention Strategies

Based on the ANN model's insights, the following targeted strategies are recommended to reduce customer churn:

1. **Personalized Onboarding for New Customers**: Provide a personalized integration experience with welcome packages, tutorials, and periodic check-ins in the first few months to increase and enhance engagement.
2. **Loyalty Rewards for Long-Term Customers**: Incentives to retain high-value customers on premium plans through exclusive discounts, rewards, and additional resources and features.
3. **Upsell Features to Basic Plan Subscribers**: Offer Basic plan customers the option to upgrade and enhance or bundle services at a discounted rate. This can improve user satisfaction and retention.
4. **Proactive Retention for High-Risk Customers**: Proactively reach out to high-churn-risk customers with personalized offers and account reviews to address any concerns.

9. Limitations and Suggested Improvements

While the model yielded insightful results, certain limitations were encountered, with proposed solutions for future development:

1. **Data Quality**:
   - **Challenge**: Incomplete or missing data on some features can affect model predictions. Although missing values must be handled through imputation.
   - **Solution**: Future iterations can improve data quality of the data by collecting more open data, such as research and surveys with customers or additional data sources, to reduce missing values.
2. **Class Imbalance**:
   - **Challenge**: There was a notable imbalance between churn and non-churn instances, which can skew model predictions.
   - **Solution**: Techniques like SMOTE (Synthetic Minority Oversampling Technique) can help balance the data. Using oversampling or undersampling methods can improve the model's accuracy in detecting churn.
3. **Model Interpretability**:
   - **Challenge**: ANN models often operate as "black boxes," making it challenging to interpret feature impacts.
   - **Solution**: Implementing Explainable IA (XAI) methods such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) may increase interpretability, Help identify the most influential features and increase and enhancing transparency.