**Assignment**

Use the "from the expert" (FTE) jupyter notebook as a starter for this assignment, and ask your instructor questions if you need help.

Use the `churn\_data.csv` file to carry out a similar EDA and visualization process as what we did in the FTE. Create at least 2 EDA plots. Write a short analysis at the end of the assignment in markdown. Interpet the plots you created.

**Data science process steps this week**

We will carry out the first two parts of the CRISP-DM data science process this week:

**1. Business understanding**

This is customer churn data for a telecommunications company. Customers can have phone as well as other services. The company is looking to reduce customer churn, where customers stop using the company's services and cancel their account. The 'Churn' column has a binary target, yes or no, that denotes if a customer churned. We want to create a machine learning model to predict the Churn target using the other available data in the dataset. Ideally, we will deploy this model to integrate with the company's database, so that a churn risk column is created for each customer. This will enable customer service reps and others to devise and use strategies to reduce churn.

**2. Data understanding**

Carry out some EDA as we did in the FTE. Create a histogram like we did in the FTE, where we plot a numeric column with the target as the 'hue'. Optional challenge: create other plots with the target as the hue, such as bar plots for the categorical columns.

**Model Evaluation Metrics**

Accuracy: 0.79

**Confusion Matrix**

[[930 106]

[183 190]]

**Classification Report**

precision recall f1-score support

0 0.84 0.90 0.87 1036

1 0.64 0.51 0.57 373

accuracy 0.79 1409

macro avg 0.74 0.70 0.72 1409

weighted avg 0.78 0.79 0.79 1409

**Interpretation**

**Accuracy -** The model achieved an accuracy of 79%, which represents the overall correct predictions out of the total predictions.

**Confusion Matrix** - The confusion matrix provides a detailed breakdown of the model's predictions:

True Negatives (TN) for class 0 (non-churn): 930

True Positives (TP) for class 1 (churn): 190

False Positives (FP): 106 (Predicted as non-churn but actually churn)

False Negatives (FN): 183 (Predicted as churn but actually non-churn)

**Classification Report**

**Precision (class 0): 0.84** - 84% of the predicted non-churn cases were correct.

**Recall (class 0): 0.90** - Out of all actual non-churn cases, the model identified 90% of them.

**F1-score (class 0): 0.87** - A balanced measure considering both precision and recall for non-churn class.

**Precision (class 1): 0.64** - 64% of the predicted churn cases were correct.

**Recall (class 1): 0.51** - Out of all actual churn cases, the model identified 51% of them.

**F1-score (class 1): 0.57** - A balanced measure considering both precision and recall for the churn class.

**Summary**

The model performs well in predicting non-churn cases (class 0) with high precision (84%) and recall (90%), as indicated by the high F1-score (0.87).

However, the model's performance for predicting churn cases (class 1) is not as strong, with lower precision (64%), recall (51%), and F1-score (0.57).

**Histogram with 'Churn' as Hue**

**Interpretation**

The histogram shows the distribution of monthly charges with different colors representing churn and no-churn instances.

There is a concentration of churn instances around lower monthly charges, suggesting that customers with lower monthly charges are more likely to churn.

**Bar Plots for Categorical Columns**

**Interpretation**

These bar plots show the distribution of churn (1) and no-churn (0) for different categorical columns.

The bar plots show the distribution of different categorical variables with respect to the target variable **Churn**. Each bar represents the count of observations for each category, and the bars are color-coded based on the 'Churn' status.

**Interpretation**

* **PhoneService\_Yes -**  The majority of customers have phone service ('1'). The churn distribution shows that customers with phone service tend to have a slightly higher churn rate compared to those without.
* **Contract\_One year and Contract\_Two year -** Customers with one- and two-year contracts ('1') have a lower churn rate compared to those with a month-to-month contract ('0').
* **PaymentMethod\_Credit card (automatic), PaymentMethod\_Electronic check, PaymentMethod\_Mailed check -** Customers with electronic check payment methods ('1') have a higher churn rate compared to credit card and mailed check methods.

**Proportions of Churn by PhoneService**

**Interpretation**

The proportions of churn are calculated for different groups based on **PhoneService\_Yes**.

**Churn Proportions by PhoneService**

* For customers with PhoneService\_Yes being True, approximately 90.9% churned (1) and 9.1% did not churn (0).
* For customers with PhoneService\_Yes being False, approximately 9.1% churned (1) and 90.9% did not churn (0).

**No Churn Proportions by PhoneService**

* For customers with PhoneService\_Yes being True, approximately 90.1% did not churn (0) and 9.9% churned (1).
* For customers with PhoneService\_Yes being False, approximately 9.9% did not churn (0) and 90.1% churned (1).

**Interpretation**

Customers with 'PhoneService\_Yes' being True have a higher churn proportion compared to those with 'PhoneService\_Yes' being False.

In both cases, whether 'PhoneService\_Yes' is True or False, the majority of customers do not churn.

The differences in churn proportions based on 'PhoneService\_Yes' suggest that this feature may have some predictive power in relation to churn. Customers with 'PhoneService\_Yes' being True seem to have a slightly higher likelihood of churning compared to those with 'PhoneService\_Yes' being False.

**Scatter Plot**

**Interpretation**

The scatter plot shows the relationship between 'MonthlyCharges' and 'TotalCharges,' with different colors indicating churn status.

It helps identify patterns and potential differences between churned and non-churned customers.

**Heatmap for Correlation Matrix**

**Interpretation**

The heatmap for the correlation matrix provides insights into the relationships between numerical features in the dataset, providing valuable insights for further analysis and model building.

**Positive Correlation** - A positive correlation between two variables means that when one variable increases, the other variable tends to increase as well. In the heatmap, positive correlations are indicated by lighter shades.

**Negative Correlation** - A negative correlation between two variables means that when one variable increases, the other variable tends to decrease. Negative correlations are indicated by darker shades in the heatmap.

**Strength of Correlation** - The intensity of the color in the heatmap reflects the strength of the correlation. Darker or lighter shades indicate stronger correlations, while a color closer to neutral (white) suggests a weaker or no correlation.

**Multicollinearity** - For binary features derived from one-hot encoding (0 or 1 values), correlation interpretation may not be as straightforward because these features are not continuous. However, high positive or negative correlations between them may indicate potential multicollinearity, which could affect the performance of certain machine learning models.

The heatmap assists in identifying patterns and relationships between numerical features, providing valuable insights for further analysis and model building.

In the context of the Churn prediction model:

* **No Churn (Class 0):** This represents customers who do not churn. In other words, it includes instances where the model predicts that a customer will not churn, and it turns out to be correct (True Negatives) or instances where the model predicts churn, but the customer actually does not churn (False Positives).
* **Churn (Class 1):** This represents customers who churn. It includes instances where the model predicts churn, and it turns out to be correct (True Positives) or instances where the model predicts no churn, but the customer actually churns (False Negatives).

So, in the context of the confusion matrix and classification report:

* **Class 0 (No Churn):** True Negatives (TN), False Positives (FP), Precision for Class 0, Recall for Class 0, and F1-score for Class 0 are relevant.
* **Class 1 (Churn):** True Positives (TP), False Negatives (FN), Precision for Class 1, Recall for Class 1, and F1-score for Class 1 are relevant.

These metrics help evaluate how well the model is performing for each class and provide insights into its strengths and weaknesses in predicting both churn and no churn instances.

# Analysis and Interpretation

## Model Evaluation Metrics

Accuracy: 0.79

Confusion Matrix

[[687 83]

[140 147]]

Classification Report

```shell

precision recall f1-score support

0 0.83 0.89 0.86 770

1 0.64 0.51 0.57 287

```

```shell

accuracy 0.79 1057

macro avg 0.73 0.70 0.71 1057

weighted avg 0.78 0.79 0.78 1057

```

\*\*Interpretation\*\*

Accuracy - The model achieved an accuracy of 79%, which represents the overall correct predictions out of the total predictions.

Confusion Matrix - The confusion matrix provides a detailed breakdown of the model's predictions

True Negatives (TN) for class 0 (non-churn): 687

True Positives (TP) for class 1 (churn): 147

False Positives (FP): 83 (Predicted as non-churn but actually churn)

False Negatives (FN): 140 (Predicted as churn but actually non-churn)

\*\*Classification Report\*\*

```shell

Precision (class 0): 0.83 - 83% of the predicted non-churn cases were correct.

Recall (class 0): 0.89 - Out of all actual non-churn cases, the model identified 89% of them.

F1-score (class 0): 0.86 - A balanced measure considering both precision and recall for non-churn class.

Precision (class 1): 0.64 - 64% of the predicted churn cases were correct.

Recall (class 1): 0.51 - Out of all actual churn cases, the model identified 51% of them.

F1-score (class 1): 0.57 - A balanced measure considering both precision and recall for the churn class.

```

The model performs well in predicting non-churn cases (class 0) with high precision (84%) and recall (90%), as indicated by the high F1-score (0.87). However, the model's performance for predicting churn cases (class 1) is not as strong, with lower precision (64%), recall (51%), and F1-score (0.57).

## Histogram with Churn as Hue (Monthly Charges)

\*\*Interpretation\*\*

The histogram shows the distribution of monthly charges with different colors representing churn and no-churn instances. There is a concentration of churn instances around lower monthly charges, suggesting that customers with lower monthly charges are more likely to churn.

## Bar Plots for Categorical Columns

These bar plots show the distribution of churn (1) and no-churn (0) for different categorical columns.

The bar plots show the distribution of different categorical variables with respect to the target variable Churn. Each bar represents the count of observations for each category, and the bars are color-coded based on the Churn status.

\*\*Interpretation\*\*

• PhoneService\_Yes - The majority of customers have phone service ('1'). The churn distribution shows that customers with phone service tend to have a slightly higher churn rate compared to those without.

• Contract\_One year and Contract\_Two year - Customers with one- and two-year contracts ('1') have a lower churn rate compared to those with a month-to-month contract ('0').

• PaymentMethod\_Credit card (automatic), PaymentMethod\_Electronic check, PaymentMethod\_Mailed check - Customers with electronic check payment methods ('1') have a higher churn rate compared to credit card and mailed check methods.

### ## Histogram with Churn as Hue (Total Charges)

This histogram depicts the distribution of total charges, with churn and no-churn instances differentiated by color. It reveals that many churn instances occur among customers with lower total charges.

### ## Boxplot for Monthly Charges by Churn Status

The boxplot illustrates the distribution of monthly charges for churned and non-churned customers. It is observed that the median monthly charges for churned customers are relatively higher than those for non-churned customers. Additionally, the boxplot indicates a wider range of charges for churned customers, suggesting higher variability in monthly charges among customers who churn.

These visualizations provide insights into the relationships between different features and the target variable (Churn). They highlight the significance of monthly charges and total charges in predicting churn, with lower charges associated with a higher likelihood of churn. The boxplot further emphasizes the difference in monthly charges between churned and non-churned customers.

### # Analysis and Interpretation

## Model Evaluation Metrics

Accuracy: 0.80

\*\*Confusion Matrix\*\*

[[1373 171]

[256 313]]

\*\*Classification Report\*\*

```shell

precision recall f1-score support

0 0.84 0.89 0.87 1544

1 0.65 0.55 0.59 569

```

```shell

accuracy 0.80 2113

macro avg 0.74 0.72 0.73 2113

weighted avg 0.79 0.80 0.79 2113

```

\*\*Interpretation\*\*

Accuracy - The model achieved an accuracy of 79%, which represents the overall correct predictions out of the total predictions.

Confusion Matrix - The confusion matrix provides a detailed breakdown of the model's predictions

True Negatives (TN) for class 0 (non-churn): 1373

True Positives (TP) for class 1 (churn): 313

False Positives (FP): 171 (Predicted as non-churn but actually churn)

False Negatives (FN): 256 (Predicted as churn but actually non-churn)

\*\*Classification Report\*\*

```shell

Precision (class 0): 0.84 - 84% of the predicted non-churn cases were correct.

Recall (class 0): 0.89 - Out of all actual non-churn cases, the model identified 89% of them.

F1-score (class 0): 0.87 - A balanced measure considering both precision and recall for non-churn class.

Precision (class 1): 0.65 - 65% of the predicted churn cases were correct.

Recall (class 1): 0.55 - Out of all actual churn cases, the model identified 55% of them.

F1-score (class 1): 0.59 - A balanced measure considering both precision and recall for the churn class.

```

The model performs well in predicting non-churn cases (class 0) with high precision (84%) and recall (89%), as indicated by the high F1-score (0.87). However, the model's performance for predicting churn cases (class 1) is not as strong, with lower precision (65%), recall (55%), and F1-score (0.59).

### ## Histogram for MonthlyCharges

### The histogram depicts the distribution of monthly charges for both churned and non-churned customers.The stacked bars represent the counts of customers at different monthly charge levels.

**\*\*Interpretation\*\***

**\*\*Monthly Charges Distribution\*\***

* + The majority of customers have monthly charges in the lower range.
  + There is a peak around lower monthly charges, indicating a significant number of customers with lower subscription costs.
  + There is a smaller peak around higher monthly charges, suggesting another segment of customers with higher subscription plans.

**\*\*Churn Distribution\*\***

* + The stacked bars show how churn is distributed across different monthly charge levels.
  + Customers with lower monthly charges seem to have a higher proportion of churn.

**\*\*Business Implication\*\***

* + Customers with lower monthly charges are more likely to churn. This could be due to dissatisfaction with lower-tier plans or competitors offering better deals.
  + The company might want to explore strategies to retain customers with lower monthly charges, such as promotions or targeted marketing.

### ##Bar Plots for Categorical Columns

### The bar plots show the distribution of churn for different categories in categorical columns (PhoneService, Contract, PaymentMethod).

**\*\*Interpretation\*\***

**\*\*PhoneService\*\***

* + Customers with PhoneService are more prevalent.
  + The bar plot shows the distribution of churn for customers with and without PhoneService. Churn is higher for customers with PhoneService.

**\*\*Contract\*\***

* + Most customers have a month-to-month contract, followed by one-year and two-year contracts.
  + Month-to-month contract customers have a higher churn rate compared to one-year and two-year contract customers. Long-term contracts are associated with lower churn.

**\*\*PaymentMethod\*\***

* + Electronic check is the most common payment method, followed by Mailed check, Bank transfer, and Credit card.
  + Electronic check customers have a higher churn rate compared to other payment methods.
  + Customers using Credit card or Bank transfer have lower churn rates.

**\*\*Business Implication\*\***

* Understanding the distribution of churn across categorical features helps identify factors that influence customer retention.
* Strategies to improve customer retention may include promoting longer-term contracts, optimizing payment methods, or addressing issues related to PhoneService.

### ## Line Graph - Churn Proportions Over Tenure

### The line graph shows the relationship between customer tenure and the churn proportion. As tenure (the time a customer has been using the service) increases, the churn proportion tends to decrease i.e., customers who have been with the service for a longer duration are less likely to churn.

### **\*\*Interpretation\*\***

* The negative trend in the line graph suggests that customer loyalty tends to increase with tenure.
* This insight can be valuable for the business, indicating that long-term customers are more likely to stay with the service.
* The company might consider implementing retention strategies for new customers to increase their loyalty over time.

### ## Pairplot of Numeric Columns with Churn Status

### The pairplot shows scatterplots for each pair of numeric columns (tenure, MonthlyCharges, TotalCharges) with the hue indicating the churn status. The diagonal shows kernel density plots for each numeric variable.

**\*\*Interpretation\*\***

* **tenure vs. tenure -** The diagonal line indicates a perfect positive correlation (as tenure increases, tenure also increases, which is obvious).
* **MonthlyCharges vs. MonthlyCharges -**  Similar to tenure, MonthlyCharges perfectly correlates with itself.
* **TotalCharges vs. TotalCharges -** TotalCharges perfectly correlates with itself.
* **tenure vs. MonthlyCharges -** No clear pattern between tenure and MonthlyCharges.
* **tenure vs. TotalCharges -** No clear pattern
* **MonthlyCharges vs. TotalCharges -** Some customers with higher MonthlyCharges also tend to have higher TotalCharges.

The pairplot provides a quick overview of the relationships between numeric variables, helping identify potential patterns. It doesn't show strong linear patterns between tenure and MonthlyCharges/TotalCharges, indicating that these variables might not be highly correlated.