ONE VOICE

# Final Project
# PRESENTATION
## TELCO CHURNING PROJECT

# One Voice Team

One Voice team consists of 5 people who has a different background education and experiences, but we have same purpose and interest to learn about Data Science.

**Aisyah Amini**
Medical Laboratory Technologist

**Devina Reyga**
Student of MBA ITB

**Jadrian Darrel**
Supply Chain Operation

**Naomi Simatupang**
Data Enthusiast

**Rinaldy Eka S.**
Information Manager

# Background

- Data Set Telco Customer Churn shows the detail data of customer starting from customer ID, gender, payment method, monthly-total charge, and the status of customer churn. The status of customer churn define the customer who is leaving the Telco Company Service.
- The purpose of this project is to generate the machine learning model that can predict the segmentation of customers who will leave the service (churn) based on available behavioral data.

# Identification of Activities

1. Identifying activities
2. Understanding the Business & Data
3. Data Preparation & Pre-processing
4. Exploratory Data Analysis
5. Feature Importance & Feature Engineering
6. Modelling
7. Evaluation and Final Model

# Understanding  Business & Data

## Understanding the Business

Customer churn in business means the number of loss customer or unsubscribe to the service of company with any reason. There are some benefits of analyzing customer churn such as opportunities to increase profit,  improve the customer experience,  knowing the target market, and improve quality of products.

# Understanding Business & Data

## Understanding the Data

- **Understanding the definition of each field or column** from the dataset. The field/column consists of some information such as :
  - Gender: Information of Female or Male customer
  - Payment Method: Information on the customer payment method used
  - Monthly Charges: Customer monthly billing information
  - Total Charges: the total bill of customers during the subscription.
- **Understanding the data type** from each field/column such as categorical and numerical
- **Understanding the content of data from each field**
  - Gender: Female/Male
  - Payment Method: Bank Transfer, Credit Card, Electronic Check, and Mailed Check
  - Monthly Charges
  - Total Charges
  - Churn: Yes and No

# Specification of Data

| Column Name | Description | Data Type | Data Fill |
|---|---|---|---|
| Gender | The gender of customer | Kategorical | Male/Female |
| Payment Method | The method of payment customer used | Kategorical | Bank Transfer, Credit Card, Electronic Check, and Mailed Check |
| Monthly Charge | The amount of charge that customer need to pay monthly | Numeric | |
| Total Charge | The total amount of charge that customer need to pay during the subscription | Numeric | Min= 18.8  Max= 8684.8 |

# DATA PREPARATION & PRE-PROCESSING

Data preparation and pre-processing are some important steps that involve transforming raw data into a format that's suitable for analysis. In general, the steps consist of cleaning, transforming, selecting, encoding, integrating, and reducing the data. This process help to improve accuracy, reliability, and interpretability of analysis result

Check Columns Data Type

Check Null Value Data

Payment Method Denomination

Remove Customer ID Column

Check Columns Data Type

Replace Number Separator

Change Number Separator

Convert Total Charges dtype

| | customerID | gender | PaymentMethod | MonthlyCharges | TotalCharges | Churn |
|---|---|---|---|---|---|---|
| Columns Type | object | object | object | float64 | object | object |

```
#function to replace number separator
def replacee(s):
    i=str(s).find(',')
    if(i>0):
        return s[:i] + '.' + s[i+1:]
    else :
        return s
```

```
#change the number separator
telco_data['TotalCharges'] = telco_data['TotalCharges'].apply(replacee)
```

```
#convert TotalCharges dtype
telco_data['TotalCharges'] = pd.to_numeric(telco_data['TotalCharges'], errors = 'coerce')
print(telco_data['TotalCharges'].dtypes)
```

```
float64
```

**Check Null Value Data**

**Exclude the Unnamed Variable**

↓

**Show Null Value & Percentage**

↓

**Drop Null Value and Show**

Dimension of the dataset (7043, 5)

| | Columns Type | Amount of Null Values | Percentage null values |
|---|---|---|---|
| gender | object | 0 | 0.0 |
| PaymentMethod | object | 0 | 0.0 |
| MonthlyCharges | float64 | 0 | 0.0 |
| TotalCharges | float64 | 11 | 0.16 |
| Churn | object | 0 | 0.0 |

Dimension of the dataset (7032, 5)

| | Columns Type | Amount of Null Values | Percentage null values |
|---|---|---|---|
| gender | object | 0 | 0.0 |
| PaymentMethod | object | 0 | 0.0 |
| MonthlyCharges | float64 | 0 | 0.0 |
| TotalCharges | float64 | 0 | 0.0 |
| Churn | object | 0 | 0.0 |

## Payment Method Denomination

```
[11] #unique element of PaymentMethod
     telco_data.PaymentMethod.unique()
```

```
array(['Electronic check', 'Mailed check', 'Bank transfer (automatic)',
       'Credit card (automatic)'], dtype=object)
```

```
#remove (automatic) from payment method
telco_data['PaymentMethod'] = telco_data['PaymentMethod'].str.replace(' (automatic)', '', regex=False)
```

## Remove Customer ID Column

```
telco_data.drop(columns = 'customerID', inplace = True)
```

# Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an activity to explore and summarize data to get insight and identify the pattern. There's some techniques used when doing EDA, such as: Data Visualization, Descriptive Statistics, Correlation Analysis, etc.

Data Visualization

↓

Response Variable

↓

Numerical Variable

↓

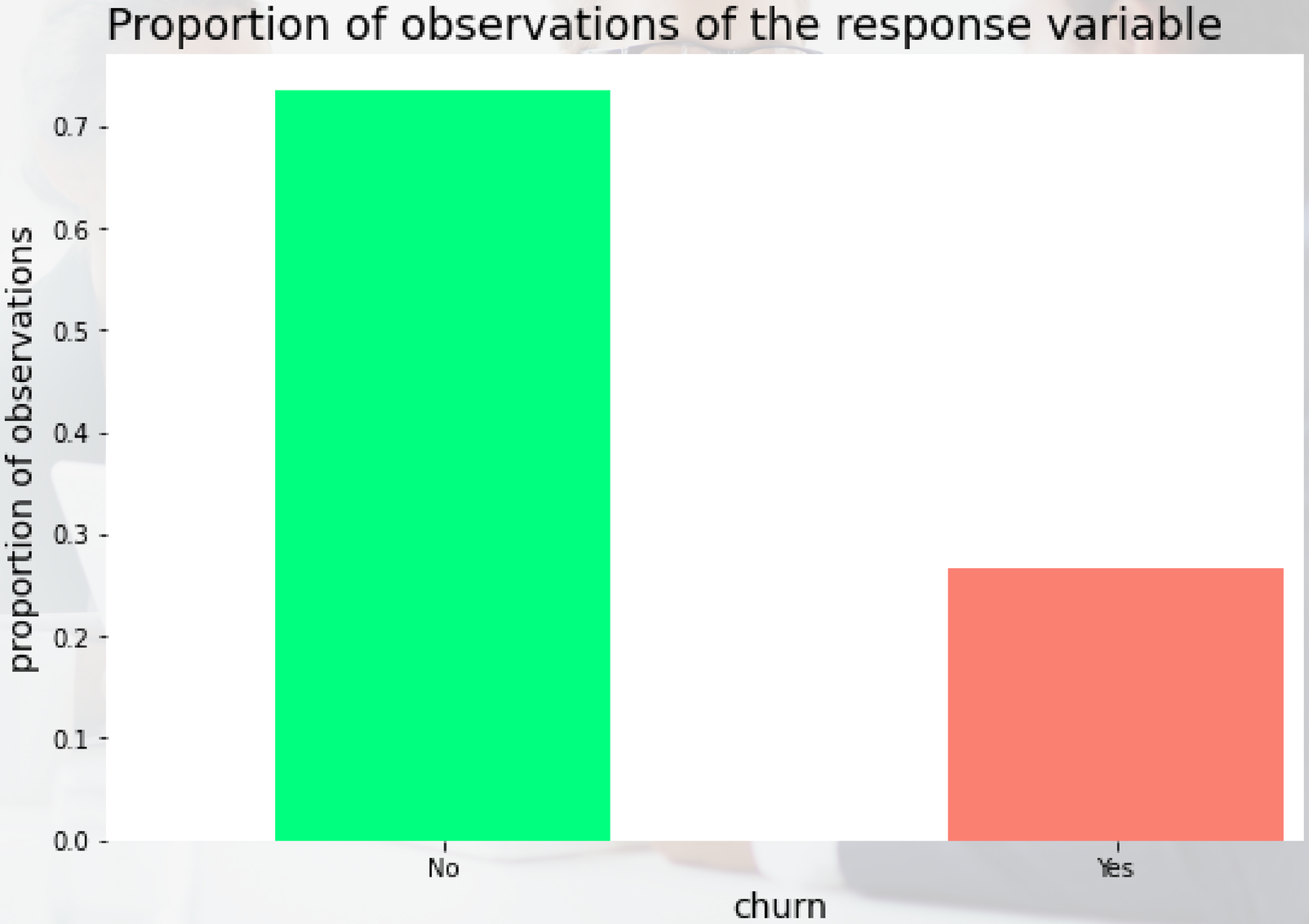
Categorical Variable

↓

Correlation
Numerical Variable

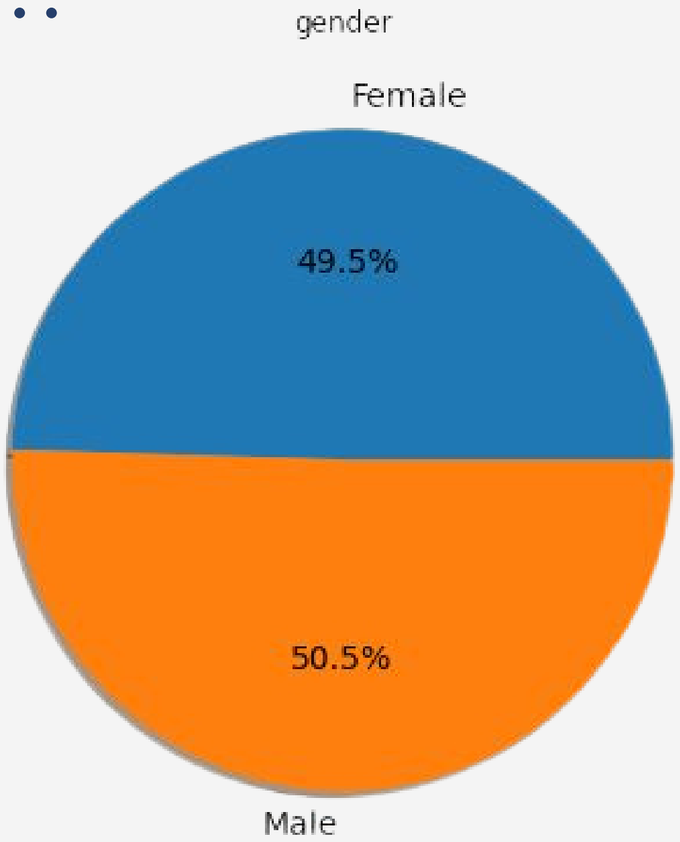
# RESPONSE VARIABLE

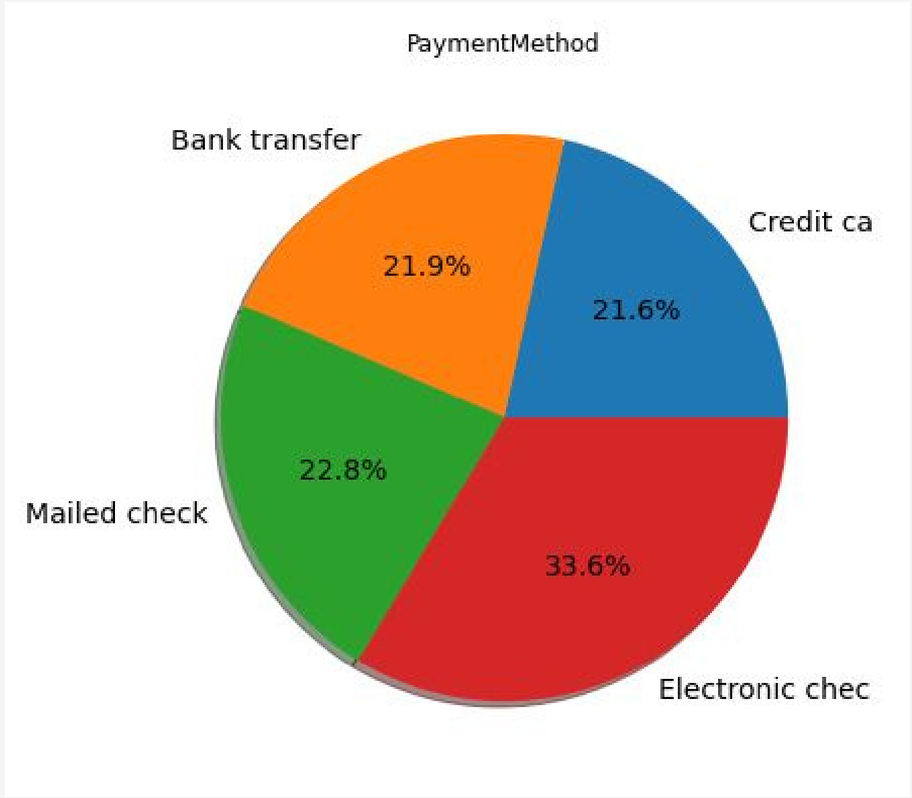
- Telco churn is a dataset consisting of 7043 rows and 5 columns.
- Identification of churn variable was obtained for 1.869 (26,58%) **YES** churn and 5.163 (73,42%) **NO** churn

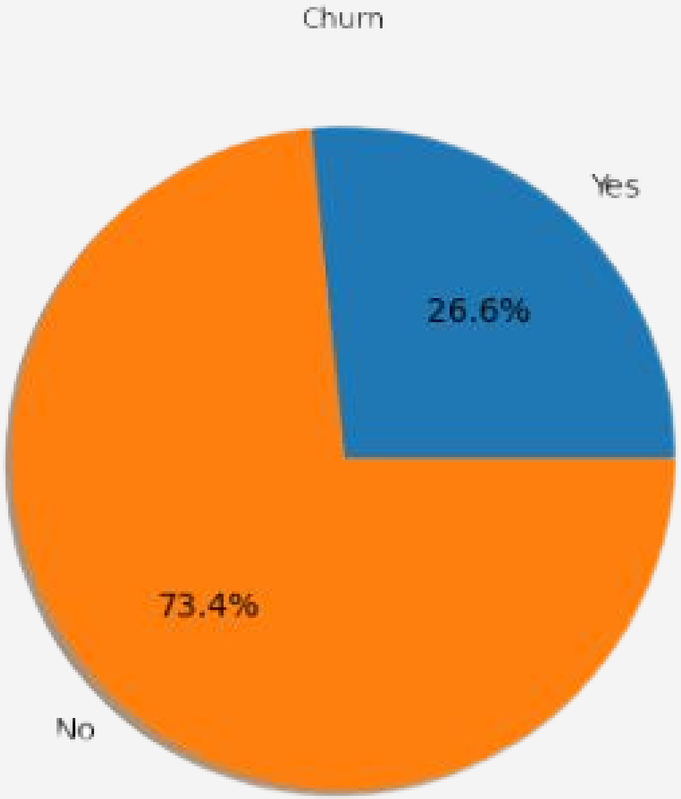


# DATA VISUALIZATION (CATEGORICAL VARIABLES)

- Percentage of gender is 49.5% for female, 50.5% for male
- Percentage of peyment method is bank transfer 21.9%, credit card 21.6%, mailed check 22.8%, and electronic check 33.6%
- Percentage of churn is 26.6% yes churn and 73.4% no churn

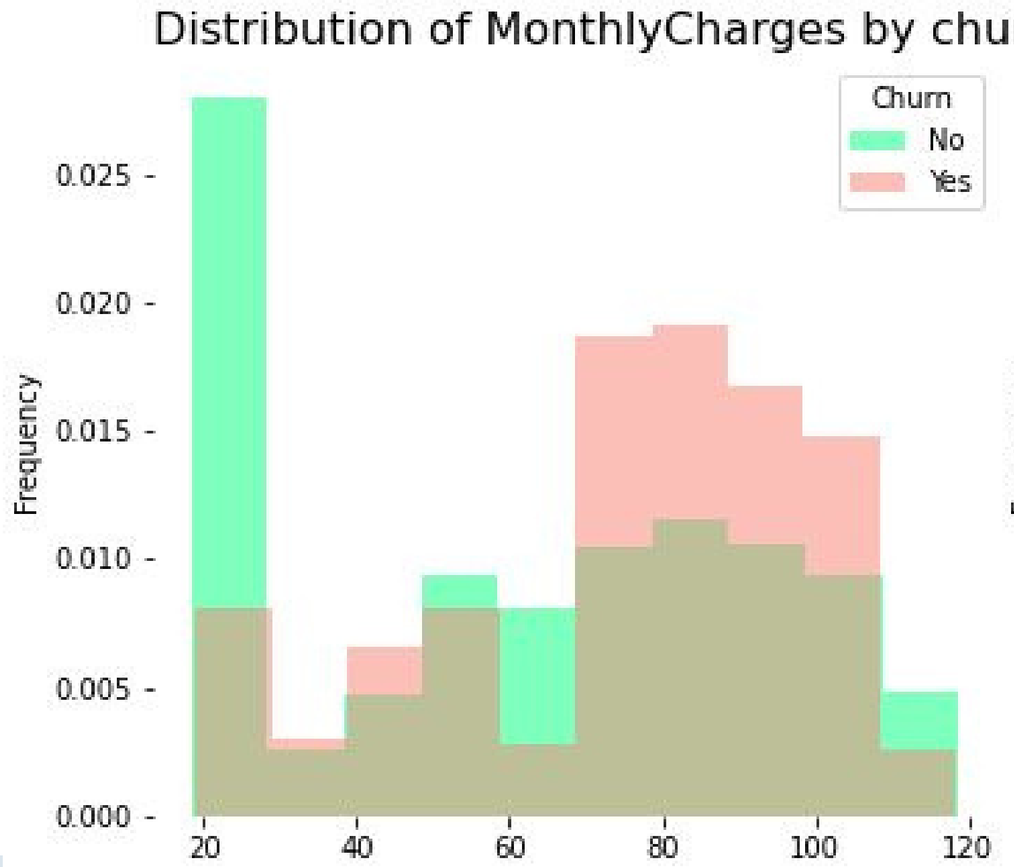


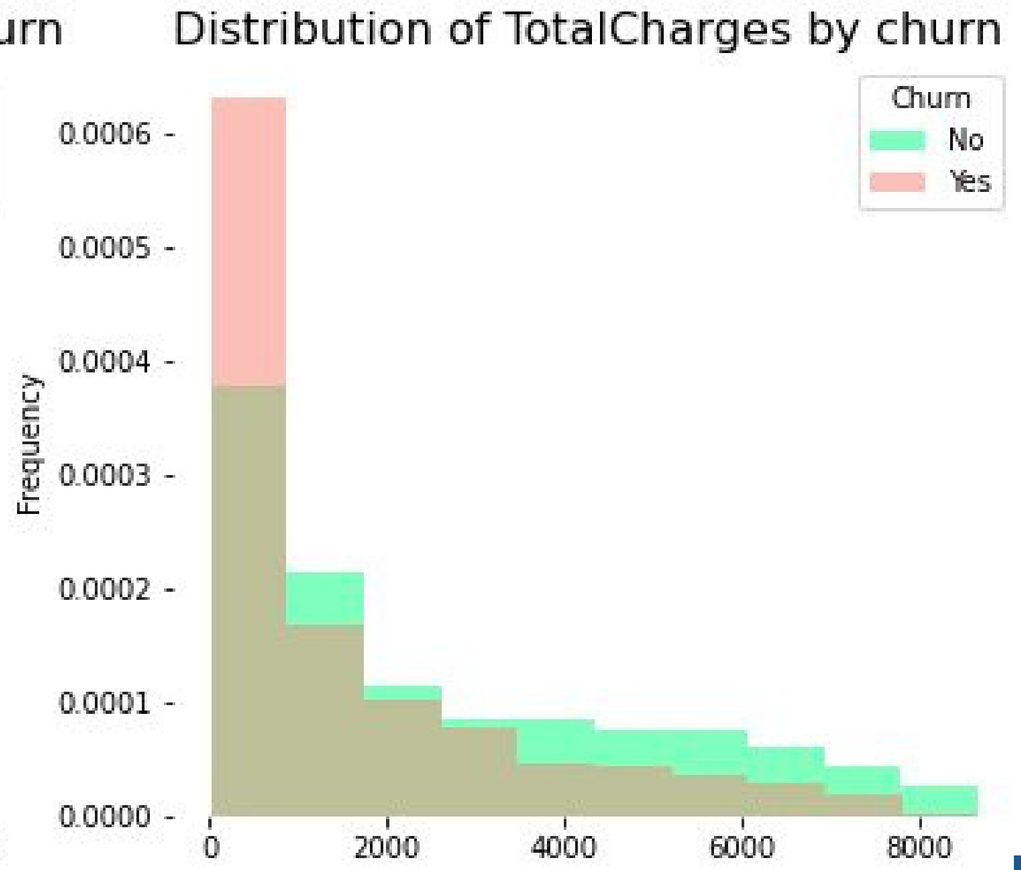



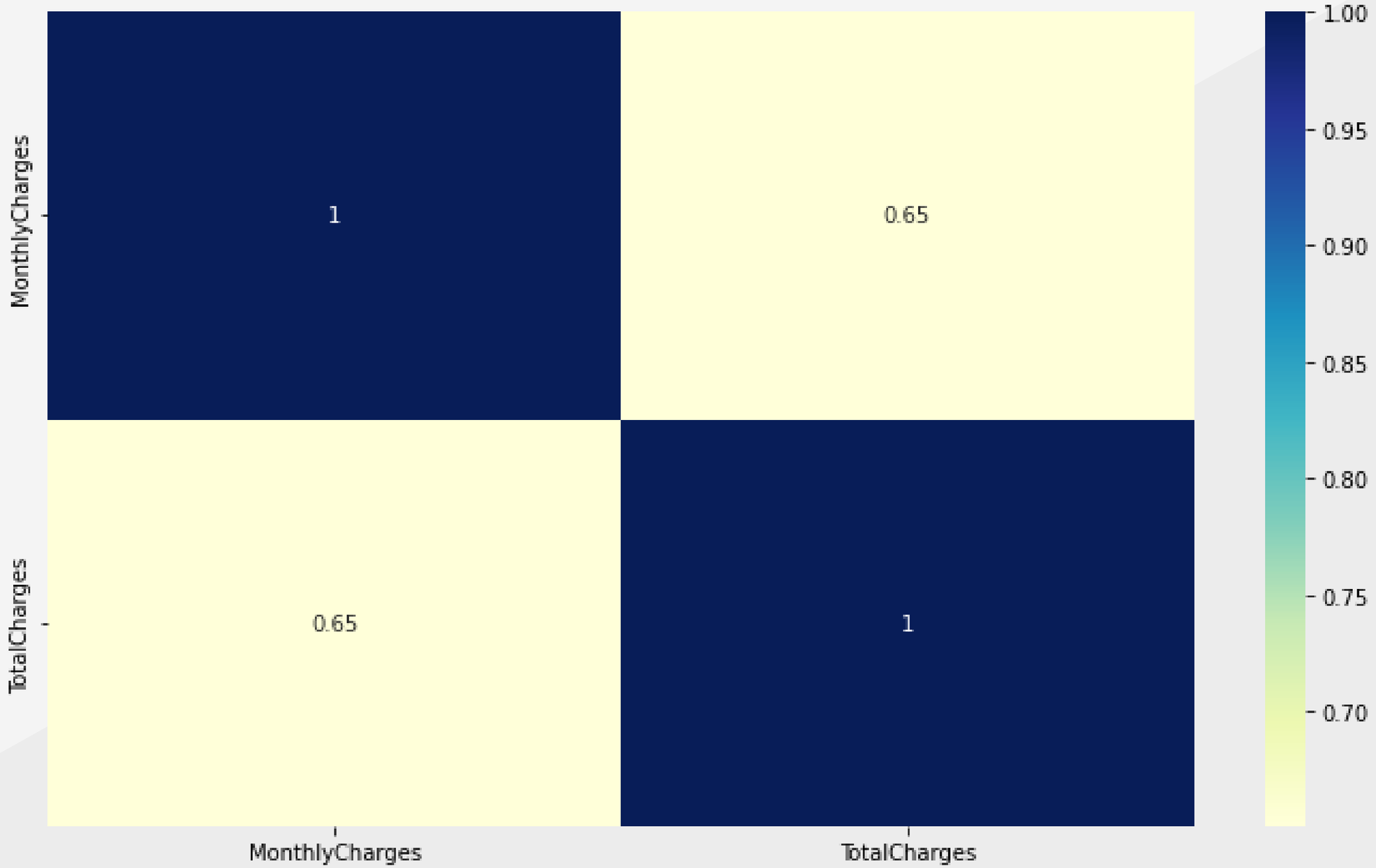
# DATA VISUALIZATION (CATEGORICAL VARIABLES)

- Distribution monthly charges is the bigger monthly charges the higher probability of churn
- Distribution total charges is customers with high total charges are less likely to churn



# CORRELATION NUMERICAL VARIABLE



# Feature Importance

- Feature importance is a technique of machine learning to identify the important variables that give the most contribution to the model. The other benefit implement the feature importance is improving the model's predictive performance, reducing overfitting, etc.
- For the Telco Case, Mutual Information is used to help understand the data and identify which variables affect the target variables in the data. If the scores owned by the payment method and gender variables are close to 0, it can be concluded that these variables do not affect the target variable.
- Therefore we can conclude that the payment method variable which has a score of 4.45% affects the results on the target variable.

```
PaymentMethod    4.44%
gender           0.00%
dtype: float64
```

# Feature Engineering

**Label Encoding** →

- To transform the categorical variable into numerical variables in machine learning. The first label encoding used for binary variables such as churn status and gender.
- The other technique used is one-hot encoding to convert payment method to be numerical variable.

| gender | MonthlyCharges | TotalCharges | Churn | PaymentMethod_Bank transfer | PaymentMethod_Credit card | PaymentMethod_Electronic check | PaymentMethod_Mailed check |
|---|---|---|---|---|---|---|---|
| 1 | 2985.00% | 2985.00% | 0 | 0 | 0 | 1 | 0 |
| 0 | 5695.00% | 188950.00% | 0 | 0 | 0 | 0 | 1 |
| 0 | 5385.00% | 10815.00% | 1 | 0 | 0 | 0 | 1 |
| 0 | 4230.00% | 184075.00% | 0 | 1 | 0 | 0 | 0 |
| 1 | 7070.00% | 15165.00% | 1 | 0 | 0 | 1 | 0 |

# Feature Engineering

**Normalization** →

- To rescale the value of numerical variables to a common scale. Normalization can improve the performance of machine learning models by reducing influenced of variable and ensuring the variables have similar scales and distribution. The normalization is used for monthly and total charges value by implementing the method of min-max.

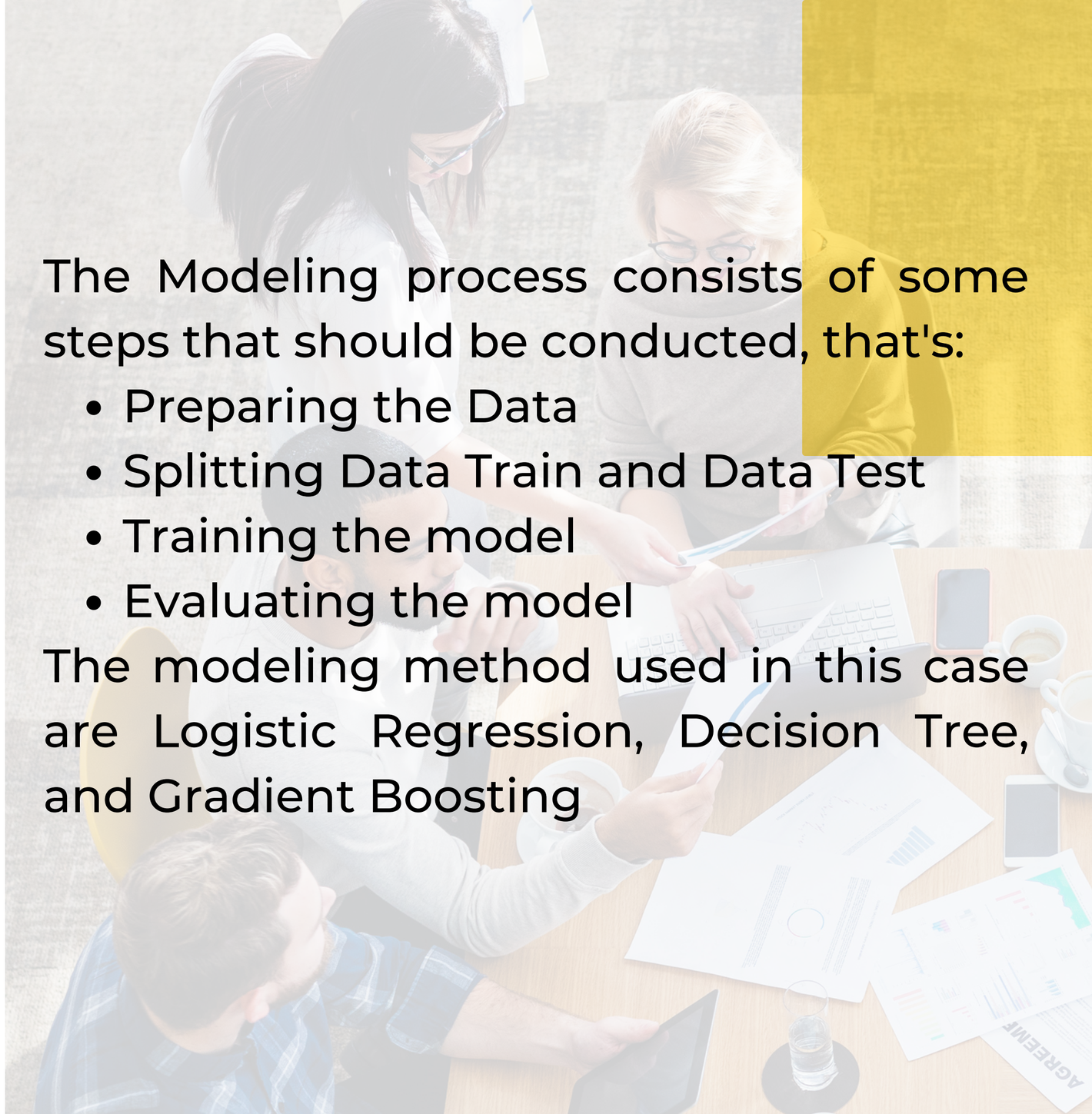
| gender | MonthlyCharges | TotalCharges | Churn | PaymentMethod_Bank transfer | PaymentMethod_Credit card | PaymentMethod_Electronic check | PaymentMethod_Mailed check |
|---|---|---|---|---|---|---|---|
| 1 | 11.54% | 0.13% | 0 | 0 | 0 | 1 | 0 |
| 0 | 38.51% | 21.59% | 0 | 0 | 0 | 0 | 1 |
| 0 | 35.42% | 1.03% | 1 | 0 | 0 | 0 | 1 |
| 0 | 23.93% | 21.02% | 0 | 1 | 0 | 0 | 0 |
| 1 | 52.19% | 1.53% | 1 | 0 | 0 | 1 | 0 |

# Modeling

Logistic Regresion

**Decision Tree**

**Gradient Boosting**



The Modeling process consists of some steps that should be conducted, that's:

- Preparing the Data
- Splitting Data Train and Data Test
- Training the model
- Evaluating the model

The modeling method used in this case are Logistic Regression, Decision Tree, and Gradient Boosting

# MODELLING RESULT

| Description | Precision | Recall | f1-Score | Accuracy |
|---|---|---|---|---|
| **Logistic Regresion** | 0.83 | 0.90 | 0.86 | 0.79 |
| **Decision Tree** | 0.83 | 0.8 | 0.81 | 0.72 |
| ***Gradient Boosting*** | 0.84 | 0.87 | 0.86 | 0.78 |

# Confusion Matrix

Positive: Churn

Negative: No-Churn

| | | Predicted Class: Positive | Predicted Class: Negative | |
|---|---|---|---|---|
| Actual Class | **Positive** | True Positive (TP) | False Negative (FN)<br>**Type II Error** | **Sensitivity**<br>$\frac{TP}{(TP + FN)}$ |
| | **Negative** | False Positive (FP)<br>**Type I Error** | True Negative (TN) | **Specificity**<br>$\frac{TN}{(TN + FP)}$ |
| | | **Precision**<br>$\frac{TP}{(TP + FP)}$ | **Negative Predictive Value**<br>$\frac{TN}{(TN + FN)}$ | **Accuracy**<br>$\frac{TP + TN}{(TP + TN + FP + FN)}$ |

# Conclusion

- Based on evaluation modeling has done **we choose the the gradient boosting method,** because the value of precision and accuracy are the greater than others modeling. So the prediction result is more precise
- We don't choose the logistic regression because the recall value is overfitting.

ONE VOICE

# THANK YOU