CUSTOMER CHURN
TELCO CUSTOMER DATA ANALYSIS

PORTFOLIO SUWARTI

MIND MAP

Customer Churn
Telco Customer Data Analysis

Use Case : Customer Churn Telco Customer Data Analysis

Periode : 2021

Use Case Summary

Objective Statement:

- Machine learning enables decision makers to predict customer churn using company-owned historical data.
- The evaluation metric that will be used is AUC (Area Under ROC Curve).

Source Data:

Dataset link

Challenges:

- Large size of data, can not maintain by excel spreadsheet
- Demography data have a lot missing values and typo

Methodology / Analytic Technique:

- Descriptive analysis
- Graph analysis
- Machine learning model

Success Criteria:

Build machine learning models that can accurately predict Customer Churn.

Expected Outcome

Machine learning models will predict whether the customer is churn or not.

SCHEME CRISP-DM

Customer Churn
Telco Customer Data Analysis

Business Understanding

Machine learning enables decision makers to predict customer churn using company-owned historical data.

Data Understanding

Data Dictionary

Column Name	Definition
customerID	Customer ID
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
PhoneService	Whether the customer has a phone service or not (Yes, No)
MultipleLines	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 an online backup or not (Yes, No, No internet service)

Column Name	Definition
DeviceProtection	Whether the customer has the device protection or not (Yes, No, No internet service)
TechSupport	Whether the customer has the tech support or not (Yes, No, No internet service)
StreamingTV	Whether the customer has TV streaming or not (Yes, No, No internet service)
StreamingMovies	Whether the customer has movie streaming or not (Yes, No, No internet service)
Contract	The customer's contract
PaperlessBilling	Whether the customer has paperless billing or not (Yes, No)
PaymentMethod	The payment method opted by the customers
MothlyCharges	The monthly charges paid by the customers
TotalCharges	The total charges paid by the customers
Churn	The customer churn status (1 - Yes, 0 - No)
DeviceProtection	Whether the customer has the device protection or not (Yes, No, No internet service)

Data preparation

Code Used:

- Python Version : 3.7.6
- Packages :
  - Pandas
  - Numpy
  - Sklearn
  - Matplotlib
  - Seaborn

Importing Dataset

Inspect The Initial Condition of Data

Data Cleansing

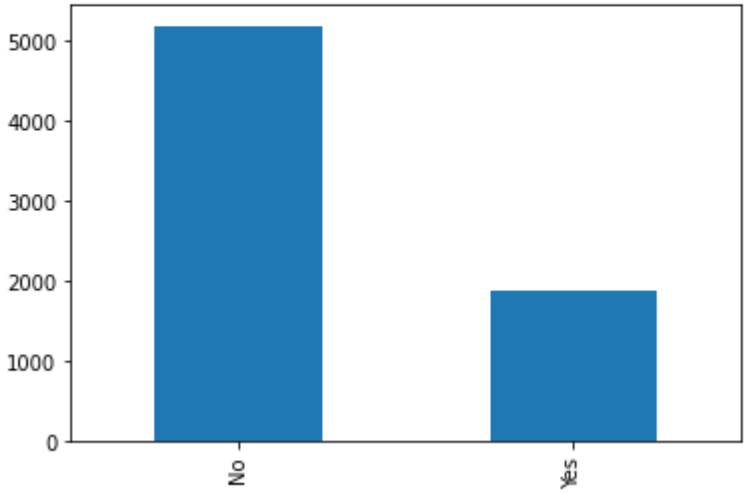
- Missing Values Checking and Handling
- Duplicates Checking
- Anomali and Outlier Detection
- Data Type Correction
- Feature Extraction

Exploratory Data Analysis

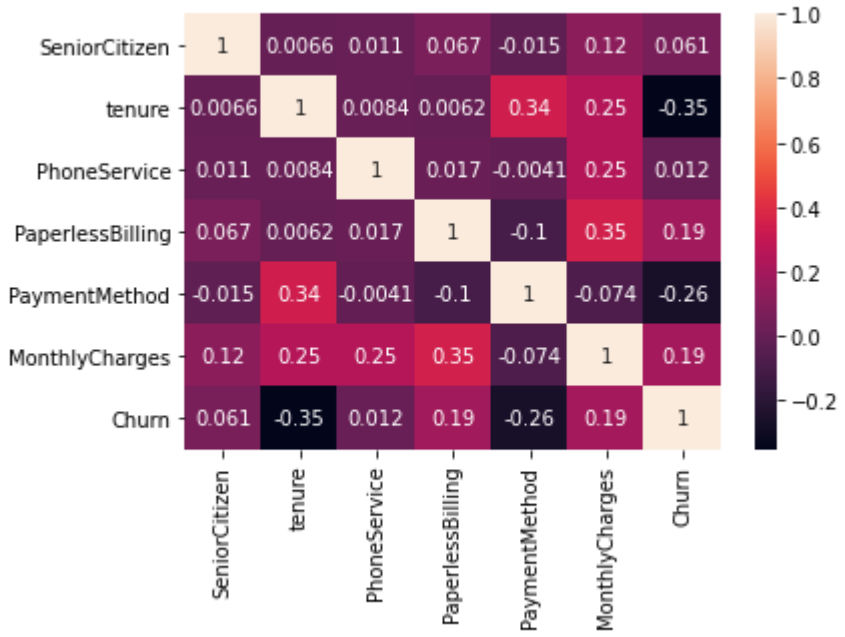
I looked at the distributions of the data and the value counts for the various categorical variables. Below are a few highlights from the pivot tables.

```
#Churn
df.Churn.value_counts().plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd64c3fe810>
```




```
sns.heatmap(df.corr(), annot=True);
```



Modeling

- Split the data into train and tests sets with a test size of 25%.
- I tried three different models.
- Evaluated them using AUC (Area Under ROC Curve).
- I chose AUC because it is relatively easy to interpret model performance.

I tried three different models:

- Support vector machine
- Decision Tree Model
- Random Forest Model

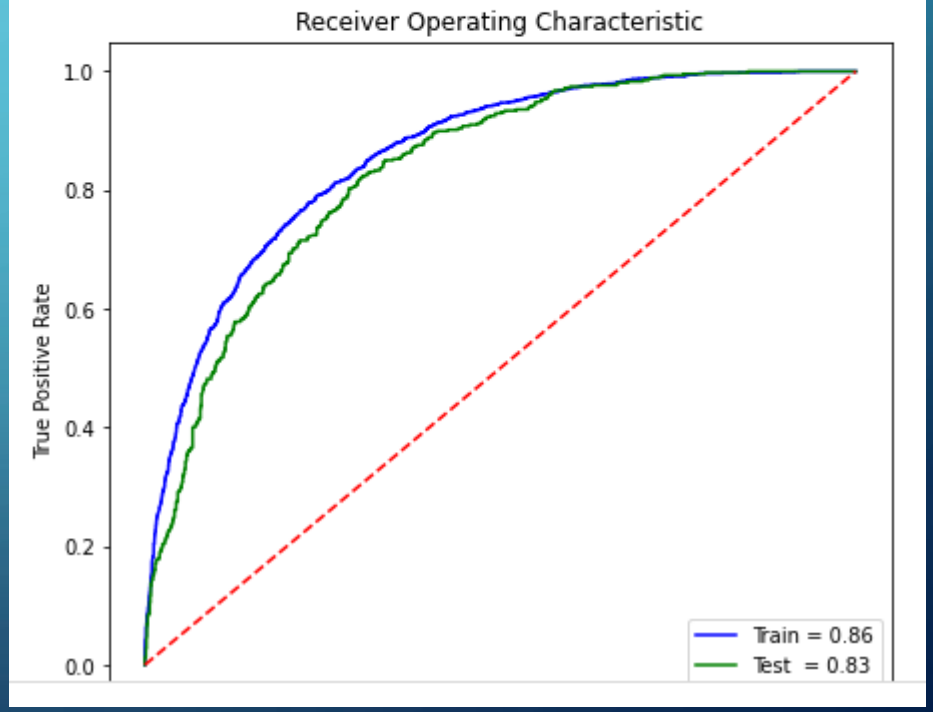
Evaluation

The evaluation metric that be used is AUC (Area Under ROC Curve).

The Random Forest model has better performence than the other approaches on the test and validation sets.

```
Area Under Curve
AUC train & test        : 85.82% & 83.28%

Confusion Matrix Evaluation
Accuracy train & test   : 81.37% & 78.88%
Recall train & test     : 47.36% & 41.54%
Specificity train & test: 93.66% & 92.35%
Precision train & test  : 72.97% & 66.21%
F1 Score train & test   : 57.44% & 51.05%
Log Loss train & test   : 6.4344 & 7.2961
```



Summary

- Machine learning enables decision makers to predict customer churn using company-owned historical data.
- The evaluation metric that be used is AUC (Area Under ROC Curve).
- I tried three different models: Support vector machine, Decision Tree Model, and Random Forest Model.
- The Random Forest model has better performence than the other approaches on the test and validation sets.

Thank You