Summary of Machine Learning Evaluation Results

Project:

“Customer Churn Prediction - Building a Model to Predict Churn in a Telco Company”

1. Data Preparation
   Problems found:
   - Missing values in 'Churn Reason'
   - Some features with incorrect data types

   Actions taken:

   - Fill in the missing value with ‘not churned’
   - Covert features into their correct data types
   - A new column called 'Churn Category' was created from 'Churn Reason' col to summarize the main causes of churners:
     - Attitude
     - Competitor
     - Dissatisfaction
     - Price
     - Not Churned
2. Exploratory Data Analysis
   Customer Attributes:
   - The occurrence of churns mostly happened among those who are:
     - not senior citizens
     - no partners
     - no dependents

   Service Subscriptions:

   - High number of churners in services:
   - Subscribe to a `home phone` service
   - Internet services:
     - Use `fiber optic` as Internet service
     - Don't use additional Internet support services such as `device protection plan` for internet equipment, `online security`, `tech support` and `online backup`
   - Entertainment services:
     - Equal churn rate between those who use and do not use `streaming TV` and `streaming movies` services
   - Subscribe to `paperless billing`

   Continuous Features:

   - 50% of the churners had stayed with the company for `less than 10 months`.
   - On a monthly basis, churners spent more than non-churners.

Others:

- Other features of the churners:
  - Most of the churners went to the `competitor` side.
  - `attitude` is also one of the factors that made them leave.
  - Mostly were `month-to-month` users.
  - A lot of them used `electronic check` as a payment method.
- Imbalance between the two classes

3. Preprocessing
   - Encode categorical features with 'Yes/No' values using a function
   - Encoding gender category
   - Encoding ordinal variable
   - Encoding the other categoric features with more than two categories
   - To ensure two or more than two independent variables are highly correlated (`Absence of Multicollinearity`)
     - Use `Variance Inflation Factor` or `VIF` to identify any significant multi-collinearity
     - Value below 5 or 10 = small collinearity
4. Model Evaluation
   - Use `Cross Validation` to compare and train different models with `default parameters`.
   - A pipeline is created to include two steps to loop through StandardScaler() and each algorithm.
   - After training the data with the selected algorithms, the one that has highest ROC-AUC score will be used as a baseline model, which is Logistic Regression.

   Hyperparameter Tuning:

   - To improve the performance of the baseline model, GridSearchCV() will be used to run across possible combinations listed in the parameter grid.
   - Pipeline is also used to include steps such as feature transformation (using StandardScaler()) and GridSearchCV().
   - After the tuning, changes have been detected as follows:
     - Accuracy increased by 0.001.
     - Precision dropped by 0.004.
     - Recall increased by 0.024 .