**Logistic Regression Assignment (SAS)**

**Dataset:** sas\_logistic\_regression\_dataset.csv  
**Target Variable:** Churn (1 = Yes, 0 = No)

**Answer the below questions. In case you need help from ChatGPT feel free to use it. But make sure you understand the AI generated SAS script.**

**Section A: Exploratory Data Analysis & Preprocessing**

1. **Import the dataset into SAS.**  
   Use PROC IMPORT or DATA step.
2. **Examine the dataset for missing values and basic statistics.**  
   Use PROC MEANS, PROC FREQ, and PROC UNIVARIATE.
3. **Visualize the distribution of Churn with respect to other variables.**  
   Use PROC SGPLOT or PROC SGSCATTER for plotting.
4. **Check correlations among continuous predictors.**  
   Use PROC CORR.
5. **Create dummy variables for Contract\_Type.**  
   Use CLASS statement or manual IF logic.

**Section B: Non-ML Approach – Classical Logistic Regression**

1. **Fit a binary logistic regression using PROC LOGISTIC with all predictors.**
2. **Interpret the coefficients and odds ratios.**  
   Use the ODDSRATIO statement.
3. **Use CLASS for Contract\_Type to fit the categorical model properly.**
4. **Evaluate model fit using:**
   * ROC Curve (PLOTS=ROC)
   * Classification table (CTABLE)
   * Concordance statistics
5. **Check for multicollinearity using VIF or correlation matrix.**
6. **Refit the model after removing statistically insignificant variables.**
7. **Use the final model to predict Churn on the original dataset.**  
   Create predicted probabilities.
8. **Classify customers based on a 0.5 cutoff and create a confusion matrix.**  
   Use DATA step and PROC FREQ.

**Section C: ML Approach – Variable Selection & Model Tuning**

1. **Split the dataset into training (70%) and testing (30%) sets using PROC SURVEYSELECT.**
2. **Apply logistic regression on the training data using different selection methods:**

* Forward
* Backward
* Stepwise  
  Use SELECTION= option in PROC LOGISTIC.

1. **Score the testing dataset using the trained model.**  
   Use SCORE statement or OUTMODEL.
2. **Evaluate model performance on test data using:**

* Confusion Matrix
* Accuracy
* Sensitivity / Specificity
* ROC Curve and AUC

1. **Compare the final ML-selected model with the classical model.**

* Which has better AUC or accuracy?
* Discuss the business implication of false positives vs false negatives.

1. **Export predicted probabilities and labels to a dataset for review.**
2. **Summarize your findings in a short paragraph:**

* Key drivers of Churn
* Most reliable model
* Any limitations or assumptions