***Supervised Learning – Practical***

***Instructions:***

* ***Maintain at least one seat distance from each other***
* ***There are two questions in this paper out of which one is on Linear regression and the other is on Logistic regression***
* ***You can choose either one of them. So, do not attempt both.***
* ***You are requested to call the invigilator once all output is ready and/or you are stuck anywhere with access or any server issue***

***You must write the data prep / model code in R or equivalent software.***

***No marks will be given without proper codes and outputs shared in excel sheets. Store outputs in separate tabs as per the steps mentioned***

***The code must run end-to-end with comments indicating the output as per the questions/ steps mentioned.***

**Linear Problem statement:**

***Data set: Linear Regression Data\_Sales***

***Dependent Variable: Sales price***

***Independent Variable: All other variable***

**Objective:** To make the best models possible from this data given the choice of dependent variable. Please note that there are some missing values and outliers present – variable imputation and treatment is entirely up to you.

1. Split the data into development (70% for building model) and validation (30% for validation)
2. Model on the development data
   1. Check correlation of the independent variables with the target variable. Store the result in an excel sheet in **CORR\_DEV** tab
   2. Choose variables with low VIF (<2.5). Show all variable VIFs and highlight the ones chosen in the **VIF\_DEV** tab

Get the final beta estimates and save the entire model output. Outputs must include – Beta estimates, R2, adj-R2, Durbin- Watson and MAPE of the model. Store results in **FINAL\_DEV** tab. (The final model should perform adequately on validation tests)

1. Measure performance on validation data
   1. Get Beta stability by recreating different samples. Store results in **Beta\_DEV\_VAL** tab
   2. Get Actual vs. Predicted, R2,adj-R2, Durbin- Watson results and MAPE of the model and store results in **VAL\_RESULTS** tab

**Logistic Problem statement:**

***Data set: Logistic Regression Data\_Responder***

***Dependent variable: Response (0/1)***

***Independent: All other variable***

**Objective:** To make the best model possible from this data.

1.       Split the data into development (70% for building model) and validation (30% for validation)

2.       Model on the development data

a.       Use an appropriate method to generate a first-cut of the model. Store all relevant outputs in excel tab: **DEV\_MODEL.**

b.      Choose variables with low VIF (<2.5). Store results in **DEV\_VIF** tab in excel highlighting the ones chosen

c.       Get the final beta estimates and save the entire model output considering validation aspects.  (The final model should perform adequately on validation tests). Store results in **DEV\_FINAL** tab**. Results include: (1) Beta estimates, (2) HL test, Actual vs. predicted, (3) KS statistic, (4) Ares under ROC.**

**d.      Show the relative significance of variables**. Store results in **DEV\_FINAL\_VARIMP tab.**

3.       Measure performance on validation data

a.       Get Beta stability by using validation data. Store results in **VAL\_FINAL\_BETASTAB**

b.      Get standard model statistics like KS, Actual vs. Predicted probability, HL Test. Store results in **VAL\_FINAL\_STATS**