DATA SCIENTIST WORK SAMPLE INSTRUCTIONS

This work sample is intended to showcase your ability to build predictive models on the datasets given. You will be required to build a logistic regression model, use the model to score a holdout dataset, and provide a write-up to some prompts describing the model results and decisions made in the process. You will be evaluated on the techniques used (appropriateness and complexity), performance of the model, and evaluation of the model results.

Step 1 – Build A Logistic Regression Model

Build a **logistic regression model** using the data in the DS_Work_Sample_Data.csv file to predict the probability of a future claim (future_clm_ind)

The dataset contains information about 60,000 new business State Farm policies:

- 40,000 observations are available for you to use in your model development. They are marked with a value of "Train" for the variable sample and have a 0/1 value for the variable future_clm_ind,
- 20,000 observations are also included as a holdout dataset. They are marked with a value of "Holdout" for the variable sample and have a missing value for the variable future_clm_ind. We will use these to test your model performance and they will need to be scored in Step 2.

Variable descriptions can be found in the DS_Work_Sample_Metadata.xlsx file.

The models will be evaluated using the area under the receiver operating characteristic curve (i.e., AUC or C-stat).

Step 2 – Generate Predictions

Use the model to obtain predictions **for all 60,000 observations** in the DS_Work_Sample_Data.csv file. The output should be the **predicted probabilities** for belonging to the positive class (future_clm_ind = 1).

Please provide predicted probabilities of a future claim in a .csv file, named **DS_Work_Sample_Scored.csv**. The submitted data must have only these 2 columns **with column headers and no row labels**:

- Policy ID (plcy_id)
- Predicted probability from the logistic regression (glm_pred)

Last updated March 9, 2023

DATA SCIENTIST WORK SAMPLE INSTRUCTIONS

Step 3 – Evaluate the Model

Complete the Model_Evaluation_Summary.docx document with responses to the prompts contained within that document. Feel free to include visualizations that are useful.

Step 4 - Submission

Please provide:

1. All the code used to create the models and other supporting analyses
    a. For Python users:
        i. If working with a .ipynb file, export the code and output to a .pdf format to submit.
        ii. If the code is in .py format, please submit a copy of the code in .txt format with screenshots of the output.
    b. For R users:
        i. If working with a .Rmd file, export the code and output to a .pdf format to submit.
        ii. If the code is in .R format, please submit a copy of the code in .txt format with screenshots of the output.
    c. For SAS users:
        i. Provide code in .sas format, and save output in .pdf format.
2. The DS_Work_Sample_Scored.csv file with the model predictions, as described in Step 2.
3. The completed Model_Evaluation_Summary.docx file, as described in Step 3.

Please include all documents in a .zip file along with a list of all attachments so we can verify that we have received everything. **Do not include your name in any of the file names or within any documents.**