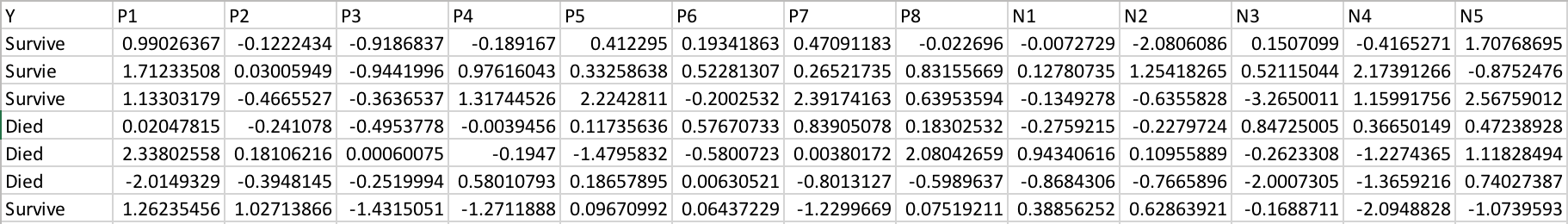
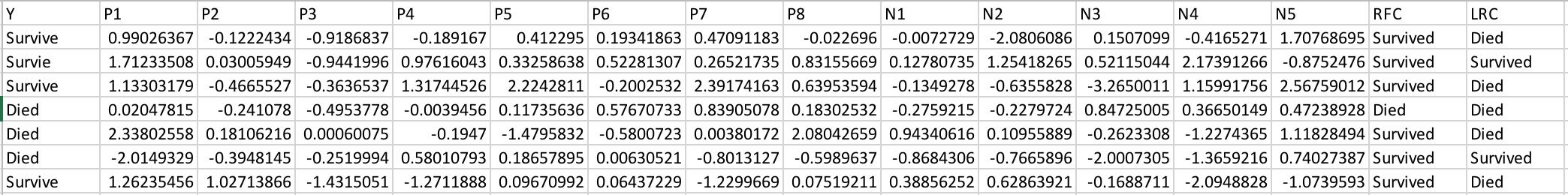
Architecture on MSDS Model Explorer App (Phase 1)

1. Generate Data (Global DataFrame “SimData”)
   1. Input
      1. Size (“Size”)
      2. Variables
         1. Predictors (“P1”,”P2”, etc)
            1. N(0, 1) (Flexible Mean and SD in Phase 2)
         2. Noise (“N1”,”N2”, etc)
            1. N(0,1) (Flexible Mean and SD in Phase 2)
      3. Response Model (“Y”)
         1. Identify Coefficients for all PXs. X = {1,2,3,….}
         2. Identify Intercept
      4. Correlation Structure (Phase 2)
   2. Output: Global Data Frame to Workspace (“SimData”)

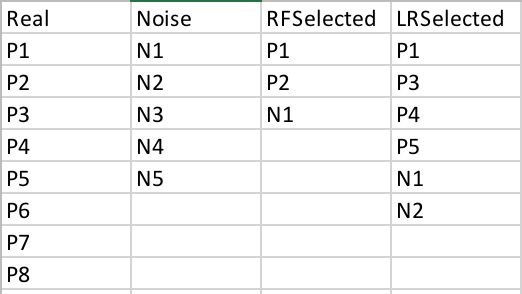


Example **SimData** DataFrame

1. Model Selection (Use “SimData” dataframe)
   1. Input
      1. Random Forrest (RF)
         1. Use mostly default hyper parameters (menus and drop downs later).
         2. Need to be able to identify response (drop down?)
         3. Need to be able to identify features (EVs) …. Default to all but Y.
      2. Logistic Regression (LR)
         1. Variable Selection (Drop down)
            1. Forward
            2. Backward
            3. Stepwise
         2. Use all other defaults
   2. Output
      1. Appended “SimData” with classifications from selected model. There will need to be an identifier that says what model the generated the classification. Example column “RFC” would be the designation of classifications that came from the RF model and column “LRC” is classifications from the LR model.
      2. Separate DataFrame called “ModelFeatures” that has the EVs/features selected from designated model. This will have a column of all real features (“Real”) from the data and a column of noise features (“Noise”) from the data. It will also have a column of selected features from each model: RF and LR: “RFSelected” and “LRSelected”.

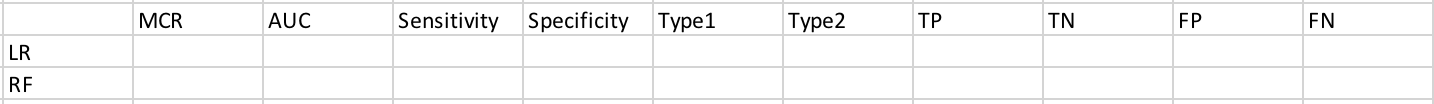


Example: Augmented **SimData** DataFrame



Example **ModelFeatures** DataFrame

1. Evaluation [Called by Model Selection Submit.]
   1. Input: “SimData” and “ModelFeatures”
   2. Output
      1. Misclassification Rate (“MCR”)
      2. AUC (“AUC”)
      3. Confusion Table
         1. True Positives (“TP”)
         2. True Negatives (“TN”)
         3. False Positives (“FP”)
         4. False Negatives (“FN”)
      4. Sensitivity (“Sensitivity”)
      5. Specificity (“Specificity”)
      6. Number of Type 1 and Type 2 Errors (“Type1”, “Type 2”)



Example **Evaluation** DataFrame