# MFM Intermediate Results

## Data Preprocessing:

* Created ”Pre” and ”Pre/Intra”(PI) data files
  + 117 and 192 variables, respectively
* Cleaned up data fields​
  + Moved [1,2] values to [0,1]​
  + Reduced precision of floats to 2 decimals.​

## Experiments ran:

* 2 Targets: transfus\_yes (Ty) and trans\_loss (Tl)
  + trans\_loss is transfus\_yes plus EBLoss > 1000
  + Ty counts - 0: 180243, 1: 5170
  + Tl counts - 0: 175069, 1: 10344
* 2 periods/stages: Prepartum and Pre-Intrapartum (combination)
* 5 Algorithms: Logistic Regression (LR), Support Vector Machines (SVC), Multi-layer Perceptron (MLP), Random Forest (RF), Gradient Boosting (GB)
  + Varying sets of hyperparameters per algorithm with dozens to hundreds of runs
* 2 sets of input features (variables):
  + All: used all 117 variables for Pre and 192 variables for Pre-Intra
  + Union50: Used Cramer to determine the top 50 trans\_loss and top 50 transfus\_yes variables and then performed a “union” or “OR” operation to produce a set of 62 variables for Pre and 66 variables for Pre-Intra.

## Initial Analysis:

* For each combination of target, period and algorithm, used 4 statistics to evaluate results:
  + Matthews Correlation Coefficient (MCC)
  + Receiver Operating Characteristic – Area Under the Curve (ROC\_AUC)
  + Precision / Recall – Area Under the Curve (PR\_AUC)
  + F2 Score – Modified F-Score skewed towards recall (F2)
* Sorted results by each of the statistics, saved these results and then selected the top hyperparameter run for each target, period and algorithm for each statistic.

## Sample Results:

###### Top runs for “Pre” period, sorted by (target, period, Alg):

Filename: pre-top\_runs.xlsx (Note: Confusion Matrix values normalized to samples per 1000)



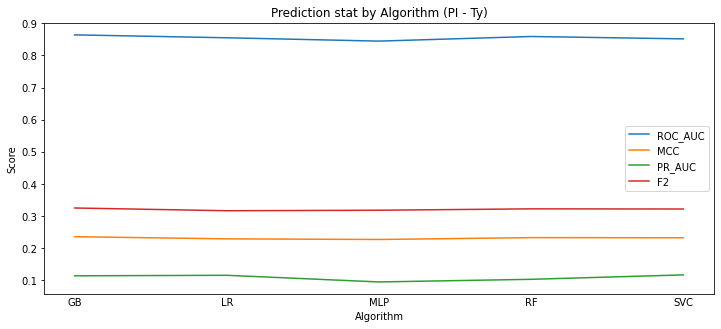
###### Top runs for “PI” period, sorted by (target, period, Alg):

Filename: pre-top\_runs.xlsx (Note: Confusion Matrix values normalized to samples per 1000)



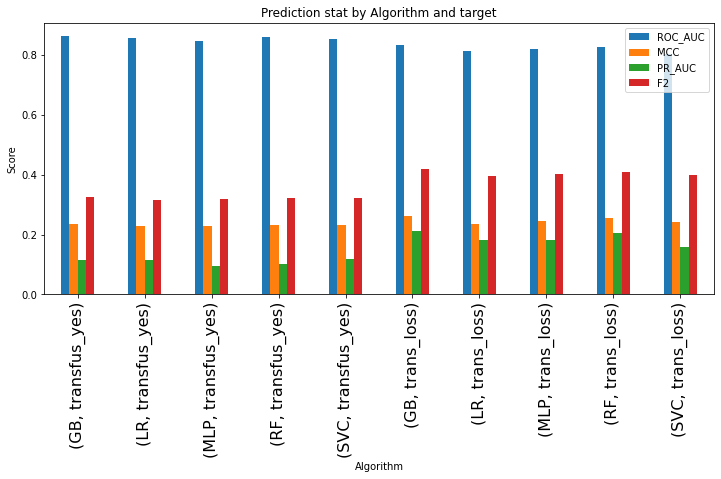
###### Prediction statistics by Algorithm (for Pre-Intra and transfus\_yes):

**(Note the consistency across the different algorithms)**

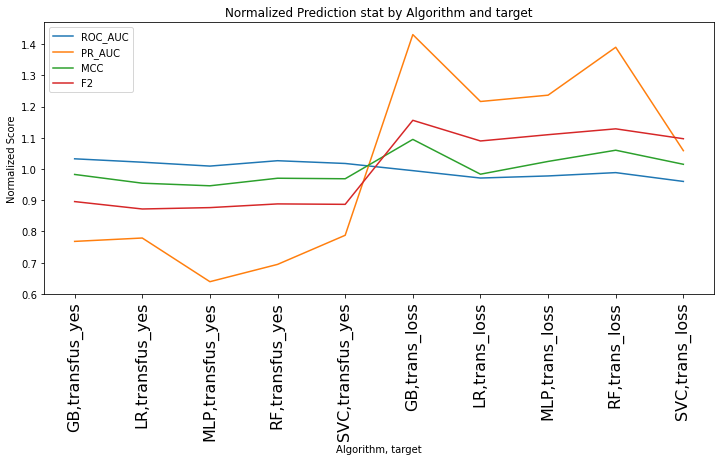
****

###### Variation of statistics based on target (trans\_loss / F2 & PR\_AUC stats)

(Note how F2 & PR\_AUC values are higher for trans\_loss than transfus\_yes)

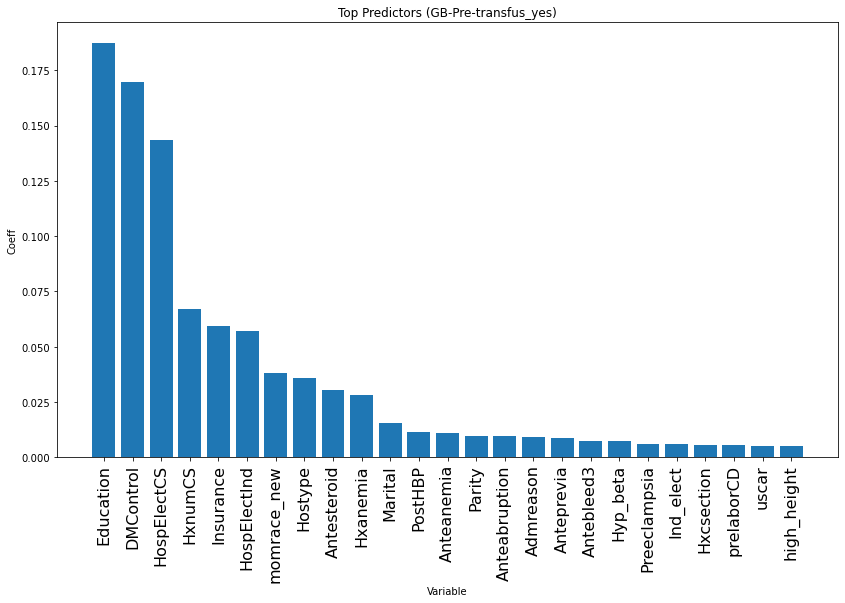


###### Same data, but statistics “normalized”

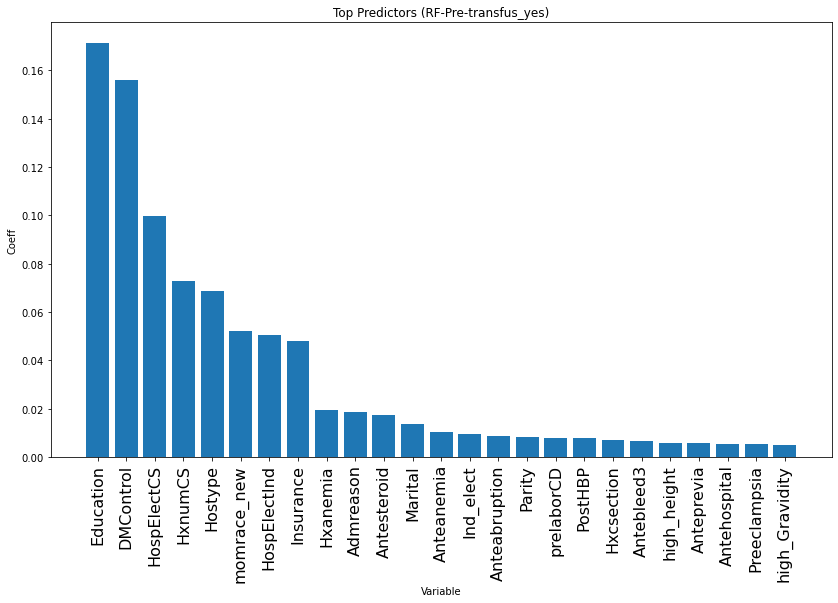


## **Top Predictors**:

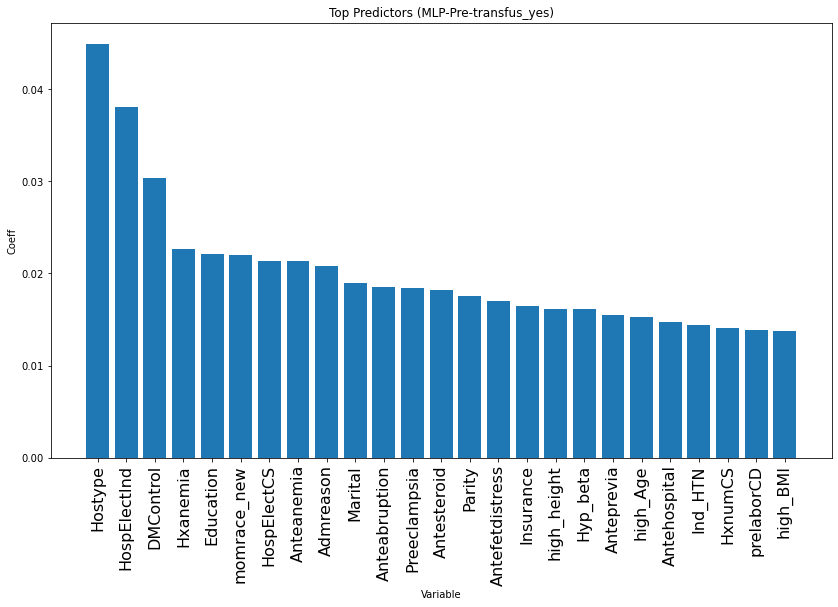
(Pre, GB, transfus\_yes)



(Pre, RF, transfus\_yes)



(Pre, MLP, transfus\_yes)



## Some Initial Observations:

* Low Precision / high False Positive rates are an issue.
* Predictions appear to be consistent across different statistics once controlled for period/stage (Pre/PI) and target (trans\_loss / transfus\_yes). Still not clear which statistic is “best”.
* Top algorithm appears to be GB followed by MLP & RF. SVC and LR seem to be consistently lower, but the chosen statistic does have an effect.
* Not surprisingly, Pre-Intra outperformed Pre given the larger number of variables, but the top results for Pre aren’t too much worse.
* At a high level, trans\_loss usually produces better predictions than transfus\_yes. However, there are exceptions based on the period/stage and the statistic used.
* Top predictor variable analysis just in the beginning stages and is still somewhat inconsistent.

## **Possible Next Steps**:

* Modify sampling method
  + Try random oversampling/bootstrapping, possibly in combination with undersampling
  + Perform k-fold cross-validation to see if it affects precision/False Negatives
  + Try a “smart” oversampling technique again on limited dataset
* Try limiting the number of features (input variables) using the Cramer coefficient.
* Try to determine how the Venkatash “CSL Hemorrhage” paper got a 0.92 ROC\_AUC using Random Forest and presumably no undersampling.
  + Why are their top predictors so different from ours? Do we have weight data?
* Other techniques possibilities include class weighting, cost-sensitive learning, expected F2