Evaluation Metrics

Model	Dataset	Accuracy	AUROC	Sensitivity (Recall / TPR)	Specificity (TNR)	F1-score
XGBoost	Training	1	1	1	1	1
	Testing	0.6600	0.6886	0.5714	0.7241	0.5854
StackingClassifier	Training	1	1	1	1	1
	Testing	**0.69**	0.6985	**0.6905**	**0.8793**	0.5373
Logistic Regression	Training	1	1	1	1	1
	Testing	0.67	0.63	0.592	0.7241	**0.6024**
RandomForest Classifier	Training	1	1	1	1	1
	Testing	0.68	**0.7005**	0.4286	0.8621	0.5294

Hyperparameters

1. Logistic Regression:
   C=100, max_iter= 100, penalty=l2 solver=lbfgs

2. RandomForest Classifier:
   max_depth= 10, min_samples_split=2,n_estimators=300

3. XGBoost:
   Subsample= 1.0, reg_lambda=1, reg_alpha=0.01, n_estimators=70,
   max_depth=10, learning_rate=0.1, gamma=0, colsample_bytree=0.8

Dataset description:

- Dataset is quite high dimensional with around 3238 columns and only few 315 rows. It can be understood through inspection the model with overfit.

1. Data preprocessing:
    - There are no duplicates in data
    - I have replaced missing values with median for logistic regression and for other I have used mean and mode. There are missing values in same column in each set of train, test and submission.
    - I handled outlier using IQR strategy i.e capped those in outside range.
    - There are inf values present in dataset. I dealt with them by replacing them with mean, median and mode.
    - Columns containing single constant values are removed as they contribute nothing in prediction
    - Trasformed using standard scaling.

2. Exploratary Data Analysis
    - It is challenge to do EDA on high dimensional dataset like this. I picked few columns randomly and plotted distplot and heatmap. Missingno is used to visualize missing values
    - To visualize outlier I randomly picked 10 column and plotted box plot. I run this code 4-5 times to get idea about outlier.
    - Used t-SNE to visualize data.

3. Feature engineering and selection
    - Target class is imbalanced (60-40). For logistic regression I have used oversampling method and for other I have used SMOTE.
    - I tried PCA, RFECV, Kbest for feature selection and PCA outformed all.

4. Cross-validation scheme
    - Passed cv as parameter while training to monitor cross validation.

5. Model Selection, Training and Hyperparameter tuning
    - Tried several models like Logistic regression, Decision Trees, Random Forest, XGBoost,Catboos, LGBM,etc. I used GridSearchCv to select the best features.

Overfitting:

The data is only few rows, which is not enough to train machine learning model that generalize well. The every model is clearly overfitting. I have used several strategies like regularization, bagging, boosting to reduce overfitting but could not achieve significat improvement.

Further Improvement:

- Inspect each column individually and do bivariate analysis to get better features.
- Extend overfitting reduction techniques.