PR2: DRUG ACTIVITY PREDICTION

Submission by: Surbhi Jain **Student Id:** 011428040
Rank: 1st **Mean F-Score:** 86.49%

Abstract
In current study, the objective was to predict the activity of drug (molecular compounds) based on provided molecular features. Along with the activity label, approximately 90 thousand binary features were provided in the raw data, which represents topological shapes and binding characteristics of the molecule. The provided sample size is relatively much smaller with respect to number of features, furthermore the classes are highly imbalanced. The ratio of dominating class to minority class is nearly 9:1. Given the dimensions of feature, the imbalance nature makes it even harder problem. To address these issues the following feature engineering was performed: 1. Dimensionality Reduction using PCA 2. Selecting K best features 3. SMOTE Sampling for balancing the dataset 4. Expanding the features to those attributes which are non-zero for positive class (Activity=1). Sometimes, too much feature engineering worsens the performance because it sometimes leads to anomalies and approximate data representation. To address that, a robust classifier is needed particularly that which is less affected to anomalies such as Ensemble based classifier. Gradient Boosting Algorithm was chosen. That gave promising result however there was opportunity for improving the model further. Therefore, an ensemble of models was used where multiple classifiers were trained with same data and the class was chosen based on majority weight for each class. The best weight given to individual classifiers for current setup was obtained using n-fold cross-validation. It is shown that the proposed setup leads to best results on the leaderboard on F-score measure. Hence, it is recommended to use this kind of ensemble of classifiers on top of highly engineered features for this kind of class-imbalance setting.

Description
One of the most important aspect of model building is data mining and feature engineering. Following steps were performed in the sequential order to build the model:

1. Data Cleaning:
   The data was provided in sparse format, where only non-zero features were provided. The data was transformed to sparse matrix format, where number of columns corresponds to all the attributes present in the training data. Attributes present in the test data and not in the training data were ignored.
2. Feature Engineering and Selection:
   As mentioned earlier, multiple features were extracted and evaluated for this project. The following features were extracted: (i) Dimensionality Reduction using PCA (ii) Selecting K best features based on Support Vector technique (iii) Features which were present in the positive class were kept.
   A total of 1000 principal components were chosen for overall features and the non-zero features for positive class. Post
3. Minority Sampling:
   SMOTE Sampling was performed to balance the classes, minority classes were up-sampled to the scale of majority class. This helped in effective training of the model, better generalization and avoiding over-fitting.
4. Model Selection:
   Sometimes, too much feature engineering worsens the performance because it can lead to anomalies and approximate data representation. To address that, a robust classifier is needed particularly that which is less affected to anomalies such as Ensemble based classifier. Apart Gradient Boosting Algorithm was chosen. That gave promising result however there was opportunity for improving the model further. Therefore, an ensemble of models was used where multiple classifiers were trained with same data and the class was chosen based on majority weight for each class. The best weight given to individual classifiers for current setup was obtained using n-fold cross-validation.

Tools and Libraries used: (1) Xgboost (2) SkLearn (3) Imblearn (4) Scipy (5) Numpy.

Experiments and Results

For fast iteration and effective turnout time, experiments were setup in the k-fold cross-validation setting, wherein given training data was split into 90-10% ratio, with former used for training and later for validation. The following experiments were performed and the results are posted on test data:

Table 1: Experiments performed and F-score on Test

Features	Model	Main Parameters	Results (F-Score)
PCA	XGB	N=1000	67%
PCA + FS	XGB	N=1000, K=500	73%
PCA + FS + SMOTE	XGB	N=1000, K=500, MS	81%
PCA + FS + SMOTE + POS_FEAT	Weighted XGB	N=1000, K=500, MS	82.4%
PCA + FS + SMOTE + POS_FEAT	Ensemble: Weighted XGB, RF, LR, GNB, SVC, MLP	N=1000, K=500, MS, weights=[5, 2, 2, 2, 2, 2]	84.21%
PCA + FS + SMOTE + POS_FEAT	Ensemble: Weighted XGB, RF, LR, GNB, SVC	N=1000, K=500, MS, weights=[5, 2, 2, 2, 2, 2] Fine tuning on probability threshold	86.49%

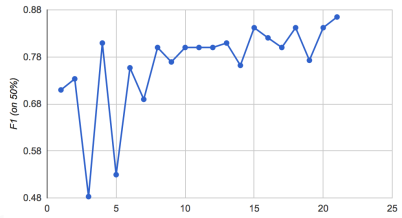
PCA: Principal Component Analysis, FS: Feature Selection, SMOTE: Minority Sampling, POS_FEAT: Adding all active features for positive class, XGB: Extreme Gradient Boosting, RF: Random Forest, LR: Logistic Regression, GNB: Gaussian Naïve Bayes, SVC: Support Vector Machine, MLP: Multi-layer Perceptron

The model performed extremely well on the validation data with Accuracy and Precision, Recall and F-Score of over 99%. The reason for extremely good results on validation set is due to the fact that the validation data was selected post feature engineering and with same SMOTE sampling, making it easy for model to differentiate.

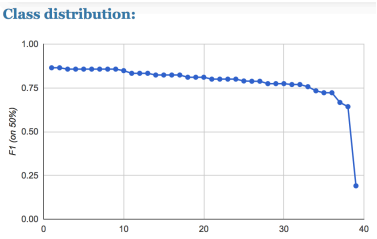
Leaderboard:

Rank	F1 (on 50%)	User ID	Submission Count
1	0.8649	11548589	15
2	0.8649	11428040	21
3	0.8571	11545378	12

Personal submissions:



Class distribution:



Conclusions, Findings and Learnings

I learnt a lot from this assignment specially feature engineering on high-dimensional data and handling imbalance data. I also learnt that ensemble techniques are extremely useful to handle anomalies and high variance in the data. Lastly, I learnt about ensemble of classifiers. It is recommended to use this kind of ensemble of classifiers on top of highly engineered features for this kind of class-imbalance setting. With this setting, the proposed model resulted in the best result on the leaderboard.