Imbalanced Mortgage Loan Default Status Classification

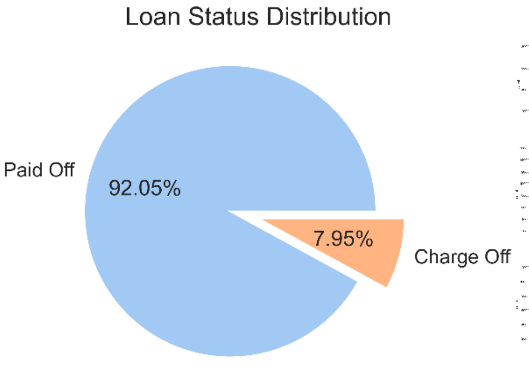
Dataset and Features

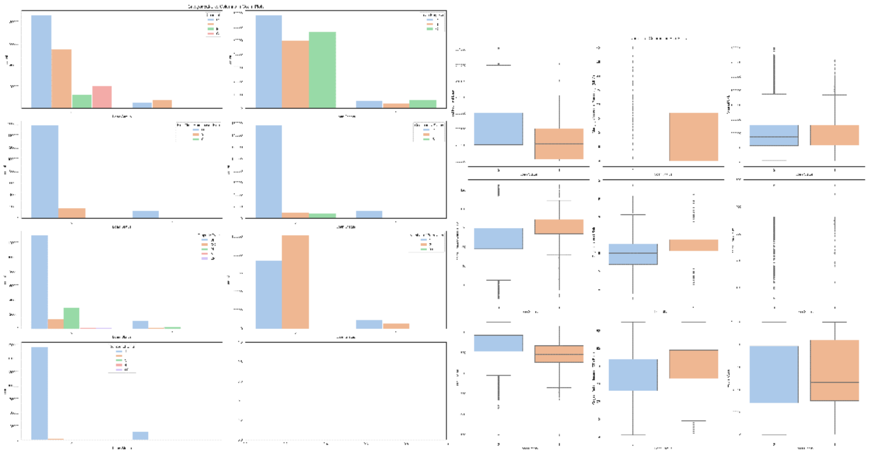
This paper applies the public Freddie Mac single-family loan level available on Freddie Mac's official website. The applied data focuses on a portion of single-family mortgage loans originated between January 1, 2007 and December 31, 2010, i.e., during and after the subprime mortgage crisis.

The dataset contains two subset datasets. The origination subset includes static mortgage loan data at inception. The monthly performance subset contains monthly dynamic performance data of each mortgage loan until the mortgage loan is terminated. A unique primary key of "Loan Sequence Number" is generated to link the to subsets and to distinguish each mortgage loan. The miscellaneous and useless information contained in the original dataset and the entries where loans were not terminated within the subset were deprecated.

The input used in this paper has 10 features. The first 8 numerical features have been selected from the origination dataset in the available 15 numerical fields characterizing the loan: credit score, original interest rate, original loan-to-value(LTV), mortgage insurance percentage(MI %), first payment year, original loan term and number of units. The other 2 categorical features have been selected from the origination dataset in the available 15 categorical fields characterizing the loan: channel, first time homebuyer flag. The categorical features are encoded with value between zero and the number of categories individually, where the not available indicator "9" being specifically encoded as zero.

The output of loan status has 2 classes with values ranging from zero to one and is selected from latest entry of the monthly performance dataset and mapped to the input corresponding to the primary key "Loan Sequence Number". The majority class of paid off loans, namely loans terminated without default, is encoded with value zero and the minority class of charge off loans, namely loans terminated in default, is encoded with value one.

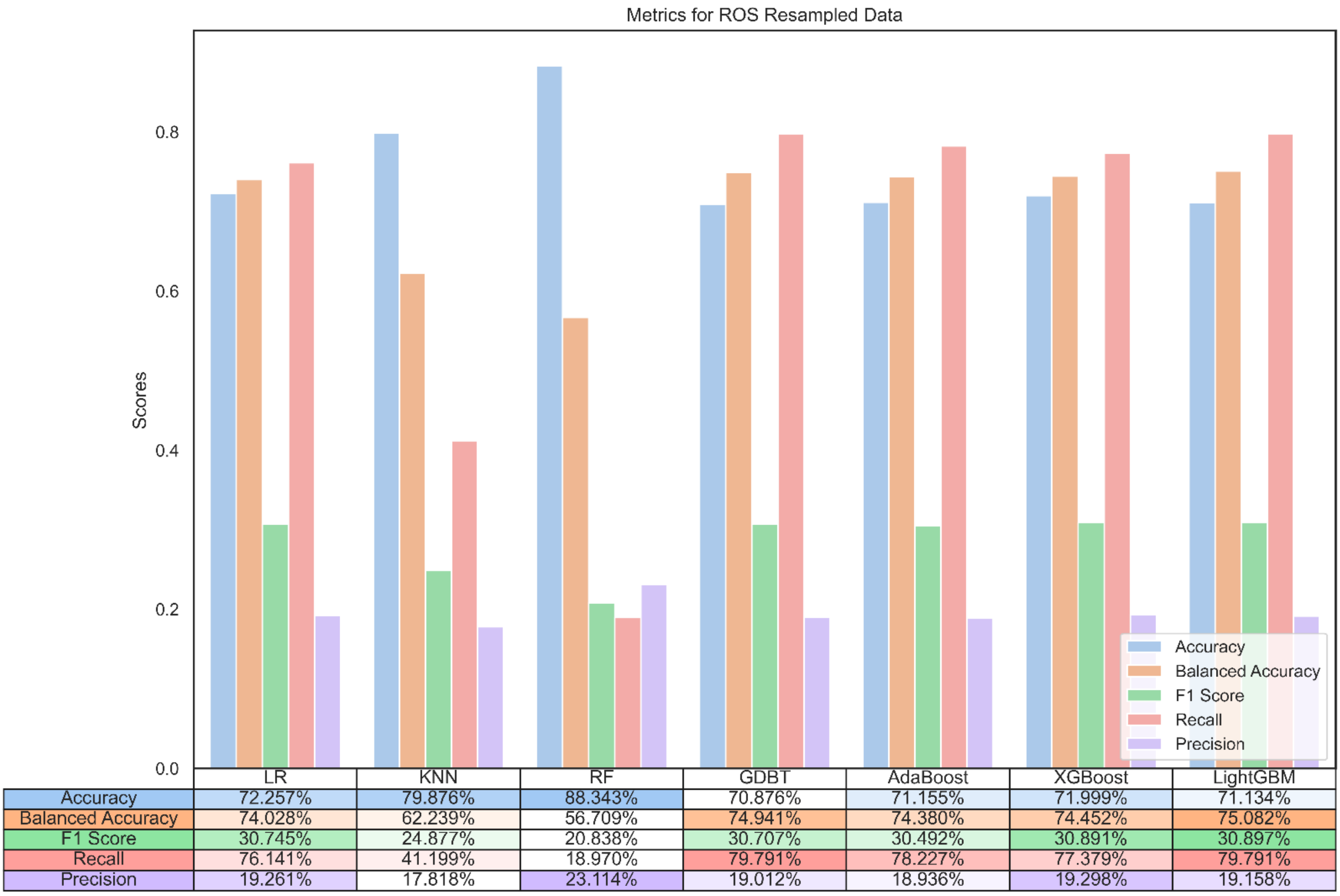




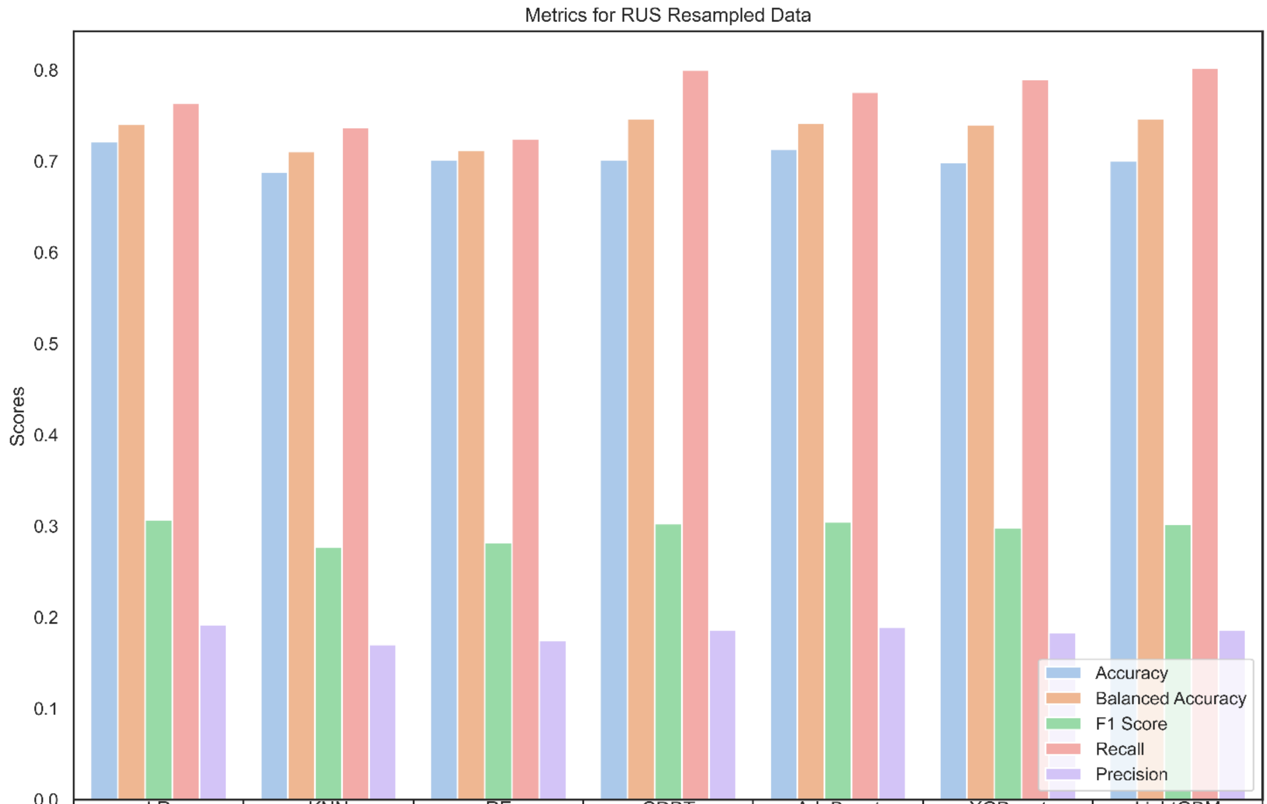
In the original dataset of over 180, 000 entries, approximately 92% of entries belongs to class 0 and 8% of which belongs to class one. Hence, the dominance of classes 0 poses the importance of resampling dataset, as model would tend to predict the majority class all the time to achieve higher accuracy. Approaches for resampling the imbalanced dataset will discussed further.

Results

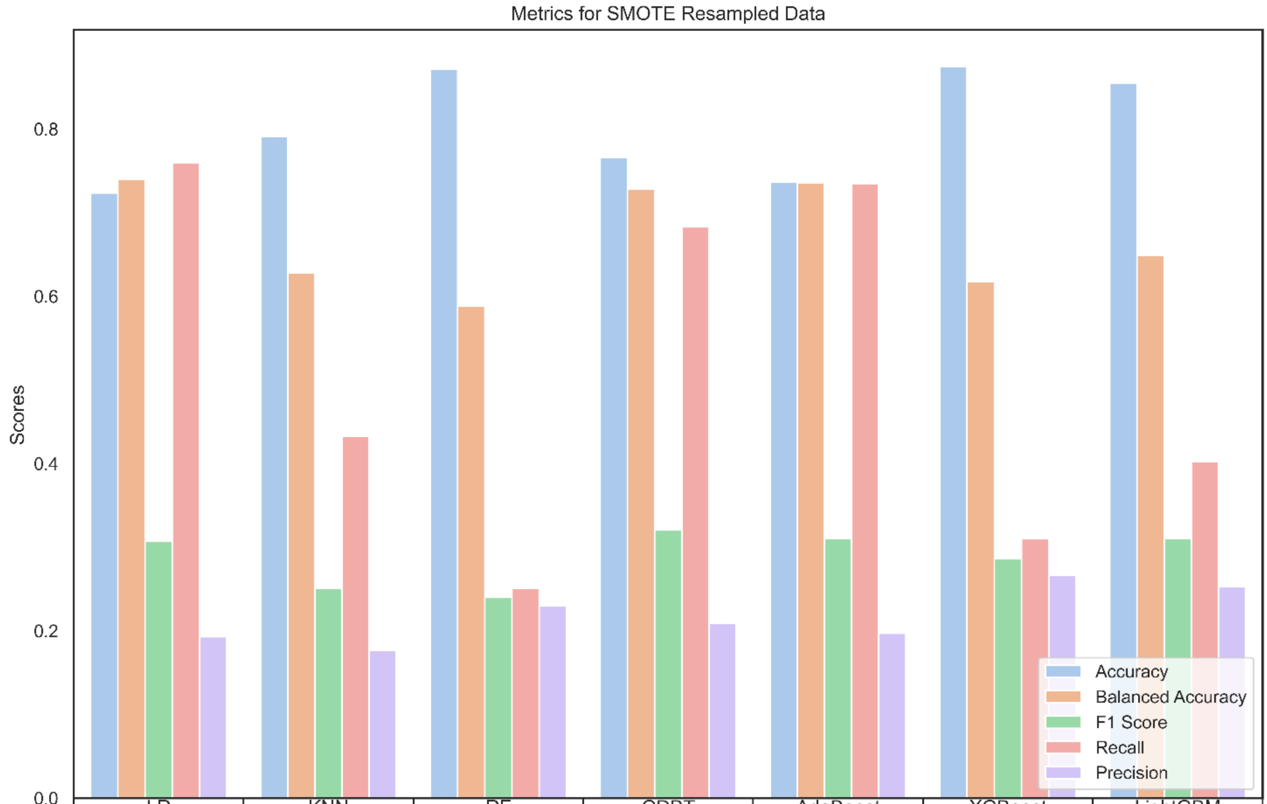
Models have been iterated over various hyper parameters to provide the best test prediction outcome. A random state of 42 has been used for all models to deliver reproducible results. Results are provided in Figures […



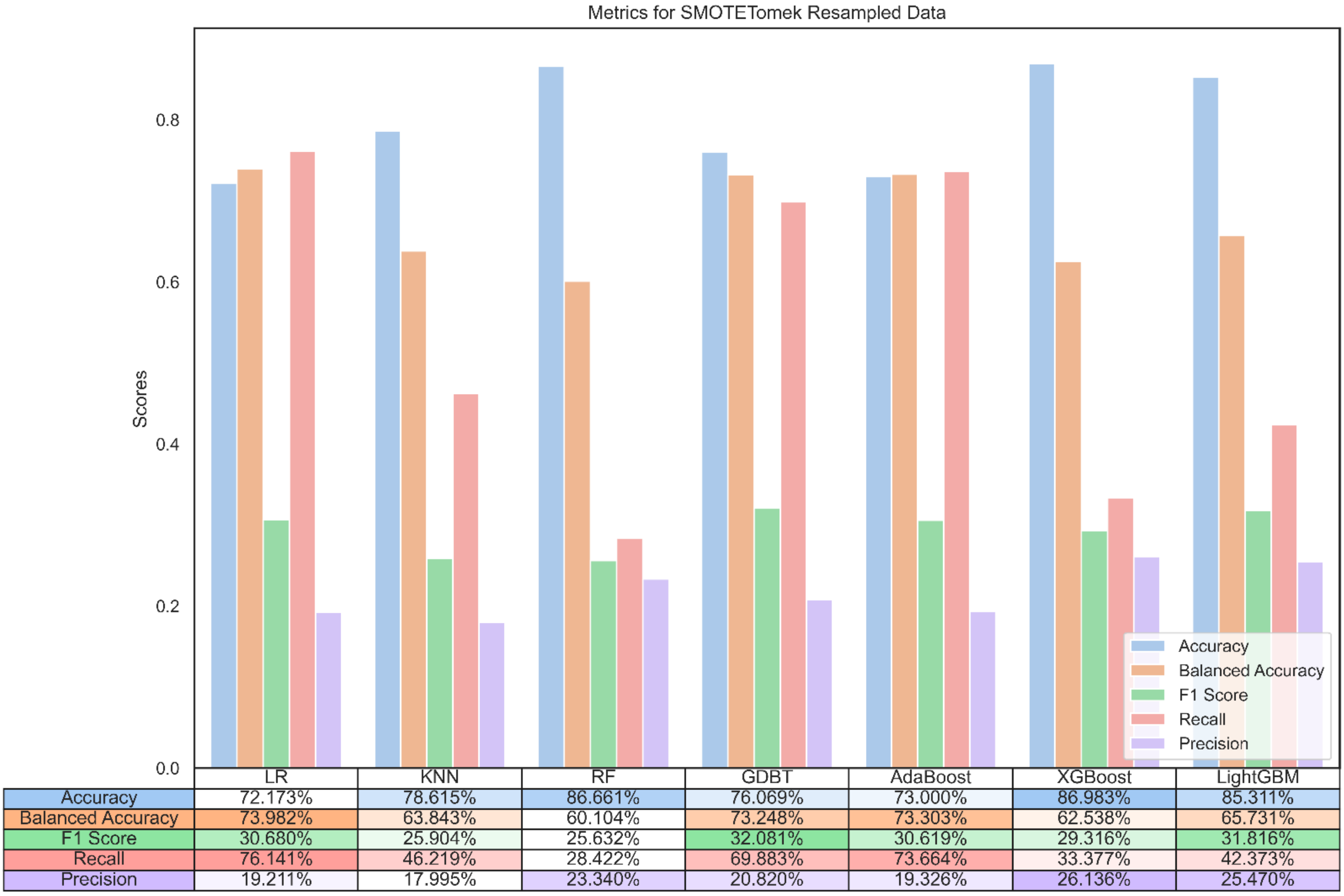
	LR	KNN	RF	GDBT	AdaBoost	XGBoost	LightGBM
Accuracy	72.257%	79.876%	88.343%	70.876%	71.155%	71.999%	71.134%
Balanced Accuracy	74.028%	62.239%	56.709%	74.941%	74.380%	74.452%	75.082%
F1 Score	30.745%	24.877%	20.838%	30.707%	30.492%	30.891%	30.897%
Recall	76.141%	41.199%	18.970%	79.791%	78.227%	77.379%	79.791%
Precision	19.261%	17.818%	23.114%	19.012%	18.936%	19.298%	19.158%



	LR	KNN	RF	GDBT	AdaBoost	XGBoost	LightGBM
Accuracy	72.136%	68.835%	70.138%	70.169%	71.345%	69.869%	70.022%
Balanced Accuracy	74.051%	71.067%	71.181%	74.646%	74.186%	74.037%	74.685%
F1 Score	30.707%	27.676%	28.177%	30.251%	30.454%	29.783%	30.216%
Recall	76.336%	73.729%	72.425%	79.987%	77.575%	79.009%	80.248%
Precision	19.219%	17.036%	17.491%	18.653%	18.946%	18.350%	18.612%



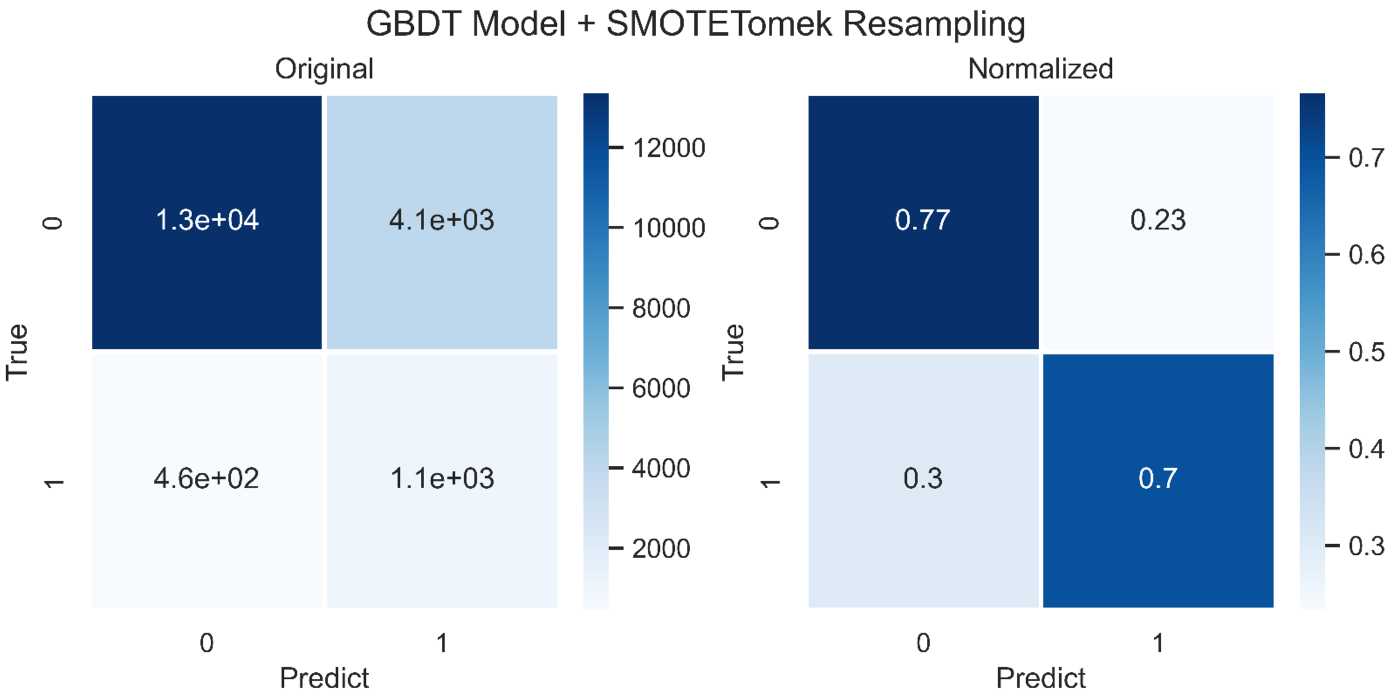
	LR	KNN	RF	GDBT	AdaBoost	XGBoost	LightGBM
Accuracy	72.373%	79.127%	87.157%	76.591%	73.660%	87.510%	85.533%
Balanced Accuracy	74.002%	62.783%	58.858%	72.819%	73.572%	61.755%	64.900%
F1 Score	30.779%	25.118%	24.017%	32.069%	31.090%	28.666%	31.055%
Recall	75.945%	43.286%	25.098%	68.318%	73.468%	31.030%	40.287%
Precision	19.301%	17.693%	23.026%	20.952%	19.717%	26.637%	25.266%



	LR	KNN	RF	GDBT	AdaBoost	XGBoost	LightGBM
Accuracy	72.173%	78.615%	86.661%	76.069%	73.000%	86.983%	85.311%
Balanced Accuracy	73.982%	63.843%	60.104%	73.248%	73.303%	62.538%	65.731%
F1 Score	30.680%	25.904%	25.632%	32.081%	30.619%	29.316%	31.816%
Recall	76.141%	46.219%	28.422%	69.883%	73.664%	33.377%	42.373%
Precision	19.211%	17.995%	23.340%	20.820%	19.326%	26.136%	25.470%

Examining the results, several observations regarding the resampling methods can be made. First of all, compare to the other resampling methods considered in this paper, random under resampling has led to a significant decrease in overall performance of all seven models. Furthermore, random over resampling has achieved mild increase in metrics other than accuracy, and significant increase in the accuracy score of model K Nearest Neighbors and Random Forests is spotted. When it comes to the SMOTE and SMOTETomek resampling methods, both resample methods have increased overall performance in various models but also ensures meaningful recall scores and precision scores. Finally, SMOTETomek has made most siginicant contributions to increase the overall performance of different metrics.

The results process a number of observations regarding the models, and the performance based on SMOTETomek resampled data is highlighted. First of all, although Random Forest, XGBoost and LightGBM has reached an ideal accuracy score over 85%, the performance varies from the abilities of detecting minority classes. For instance, Random Forest has reached a low f1 score of roughly 25% compared with 29% and 31% reached by XGBoost and LightGBM invividually. which represents low performance in detecting the minority classes. The next observation is that Gradient Decision Boosting Tree has reached highest f1 score of 32%, where 77% of majority classes and 70% of minority classes were detected properly as shown in the Figure, delivering balanced desired performance. When it comes to the evaluation between Logistic Regression, Gradient Decision Boosting Tree and AdaBoost, minor difference was spotted, thus leading to the fact that the combination of weak learners implemented by AdaBoost did not significantly increase the performance of base estimators. Finally, KNN and Random Forest have weakest performance regarding the detection of minority classes.


GBDT Model + SMOTETomek Resampling
Original
Normalized
True
Predict
0
1
1.3e+04
4.1e+03
4.6e+02
1.1e+03
12000
10000
8000
6000
4000
2000
0.77
0.23
0.3
0.7
0.7
0.6
0.5
0.4
0.3