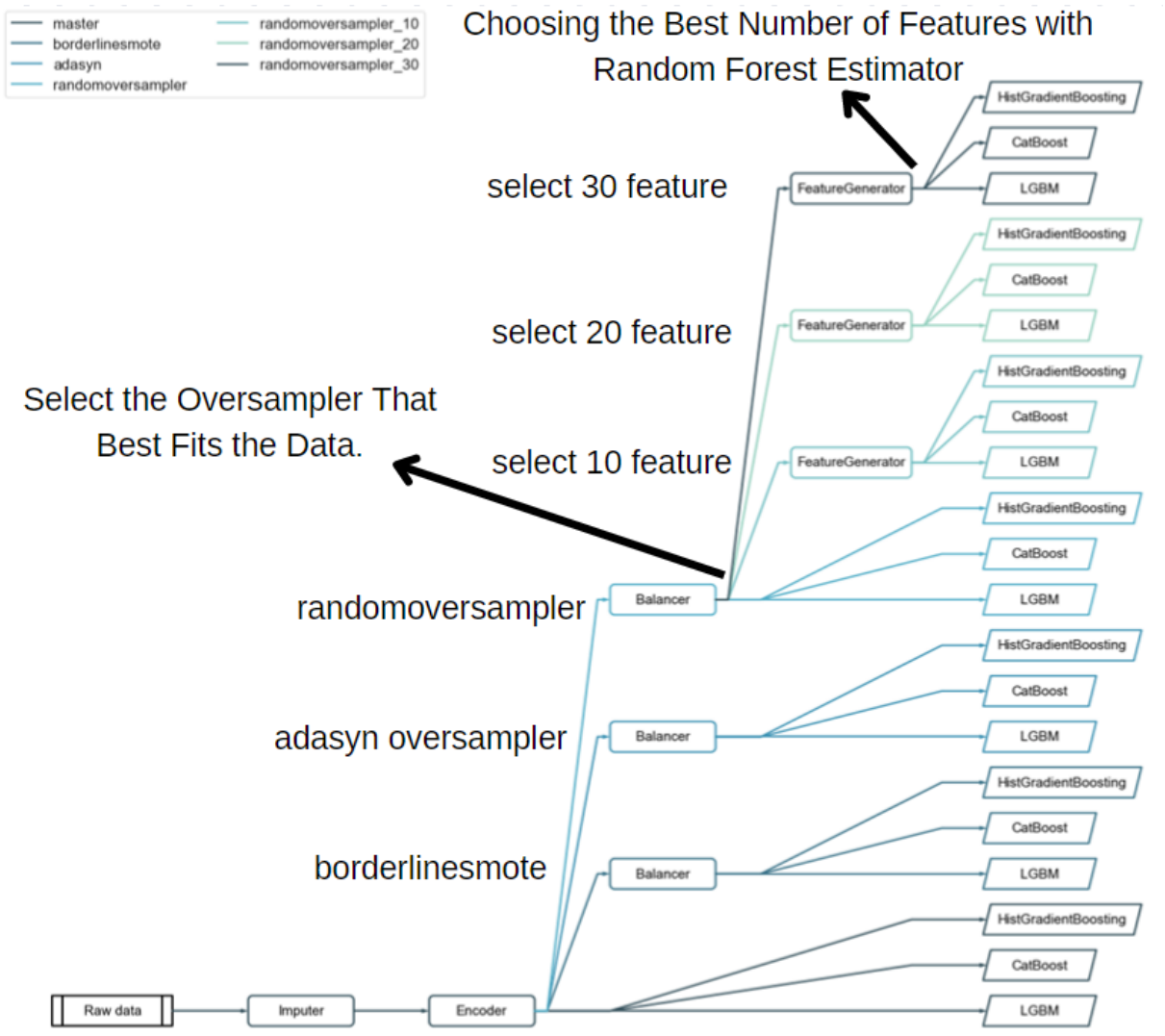
MODEL SELECTİON – PARAMETER OPTİMİZATİON

CROSS VALİDATED MODEL PERFORMANCES WİTH DİFFERENT OVERSAMPLİNG ALGORİTHMS

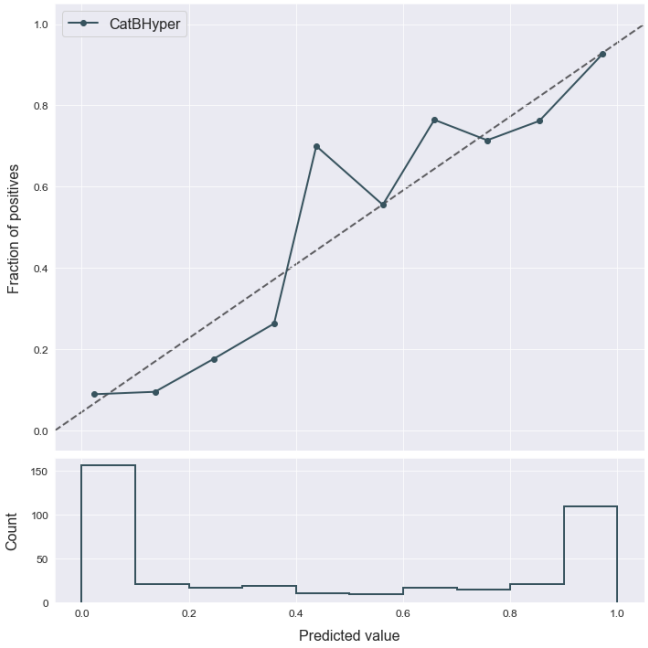
	accuracy	average_precision	balanced_accuracy	f1	jaccard	matthews_corrcoef	precision	recall	roc_auc
CatBkmeanssmote	0.873418	0.928159	0.868851	0.853801	0.744898	0.743859	0.884848	0.824859	0.926346
hGBMkmeanssmote	0.868354	0.912973	0.863202	0.847059	0.734694	0.733762	0.883436	0.813559	0.914036
hGBMbase	0.865823	0.911308	0.860908	0.844575	0.730964	0.728473	0.878049	0.813559	0.915565
CatBsvmsmote	0.863291	0.920168	0.858615	0.842105	0.727273	0.723214	0.872727	0.813559	0.922874
XGBsvmsmote	0.850633	0.917660	0.850866	0.836565	0.719048	0.699577	0.820652	0.853107	0.921319
hGBMsvmsmote	0.853165	0.913175	0.850503	0.834286	0.715686	0.702646	0.843931	0.824859	0.915954
hGBMsmote	0.850633	0.913618	0.848209	0.831909	0.712195	0.697613	0.839080	0.824859	0.917742
XGBkmeanssmote	0.853165	0.914068	0.848378	0.830409	0.710000	0.702569	0.860606	0.802260	0.916109
CatBbase	0.855696	0.908888	0.848015	0.827795	0.706186	0.709698	0.889610	0.774011	0.911393
CatBrandomoversampler	0.855696	0.915552	0.847484	0.826748	0.704663	0.710320	0.894737	0.768362	0.916680

SELECTİNG OVERSAMPLER AND DİFFERENT NUMBER OF FEATURES WİTH BRANCHES FOR OPTİMİZİNG MODEL PERFORMANCE

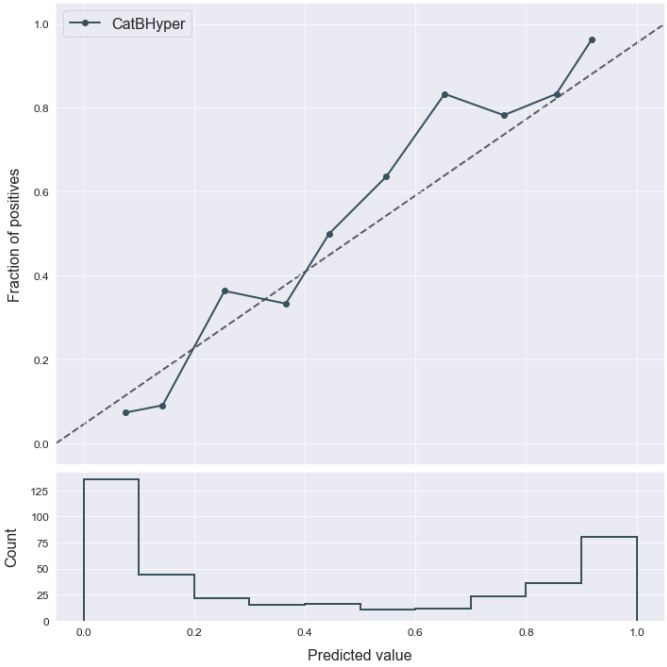


Affect of the Cross Validation on the Positive Rate with CatBoost Model

Cross-Validation



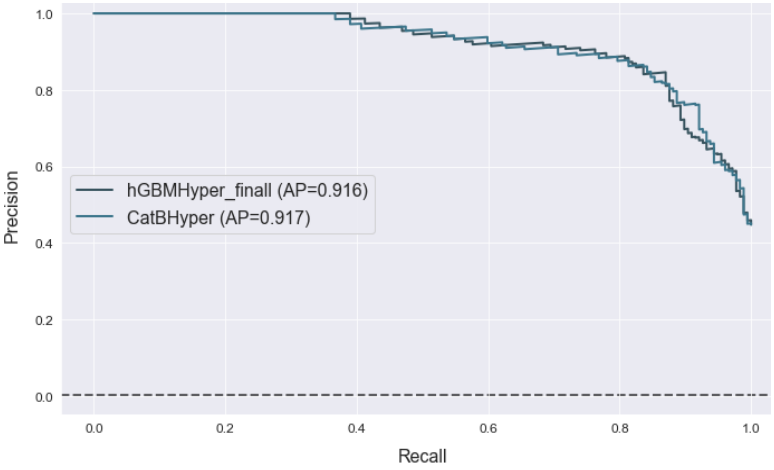
After Cross Validation



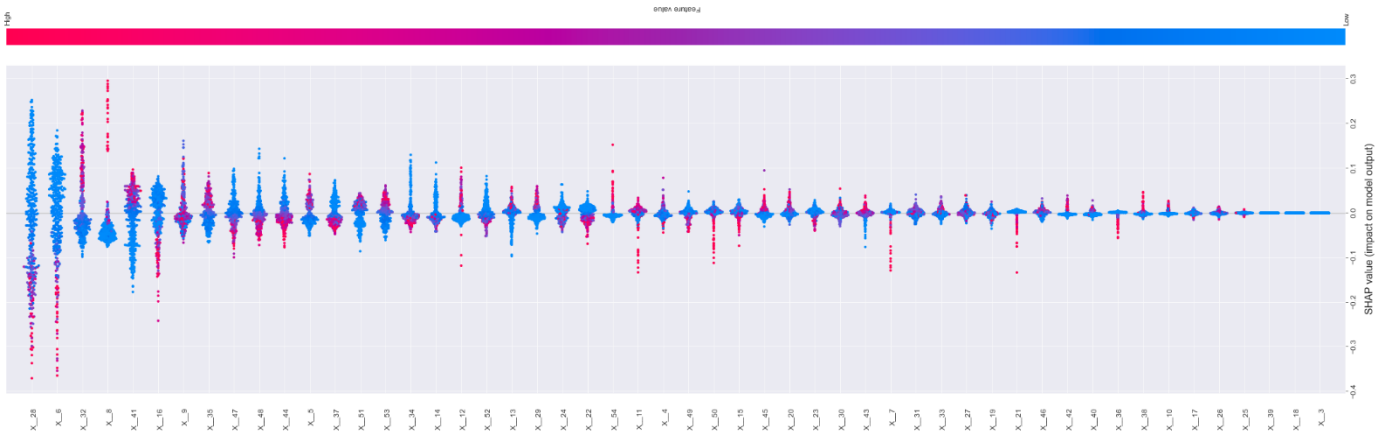
1- A model with **high recall** succeeds well in finding all the positive cases in the data, even though they may also wrongly identify some negative cases as positive cases.

2- A model with **low recall** is not able to find all (or a large part) of the positive cases in the data.

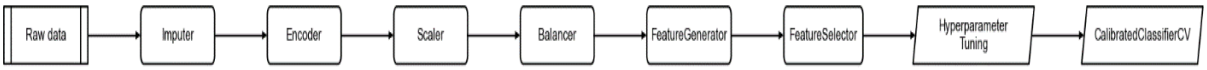
3- If you increase precision, it will reduce recall, and vice versa. This is called the **precision/recall tradeoff.**



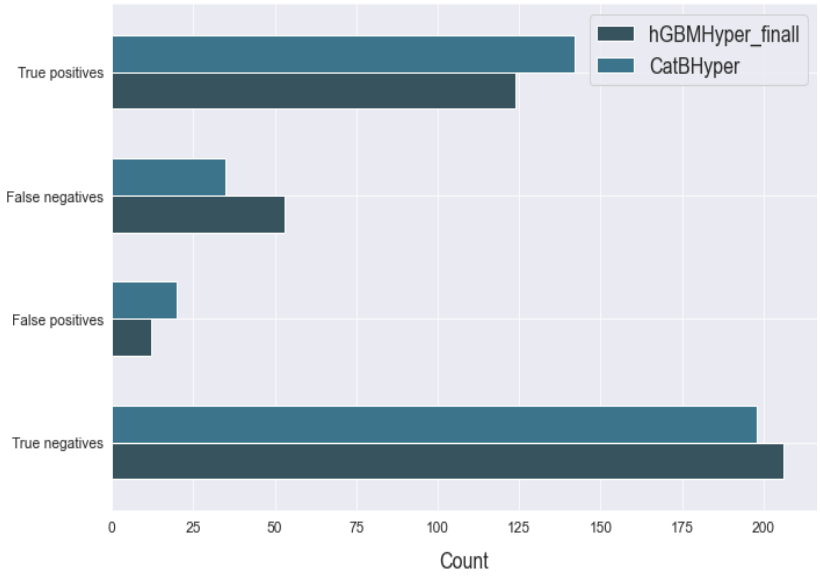
SHAP Feature Importance



CATBOOST MODEL PİPELİNE



Evaluating TP-FP TN-FN Rates of the Models



An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

CatBoost with svmsmote oversampler – hGBM with kmeanssmote oversampler