User Churn Project | Device Dependency: Hypothesis Testing

After modeling churning with a logistic regression (LG) algorithm, here we explore tree based models with aggregation: random forest (RF) and gradient boost machine (GBM).

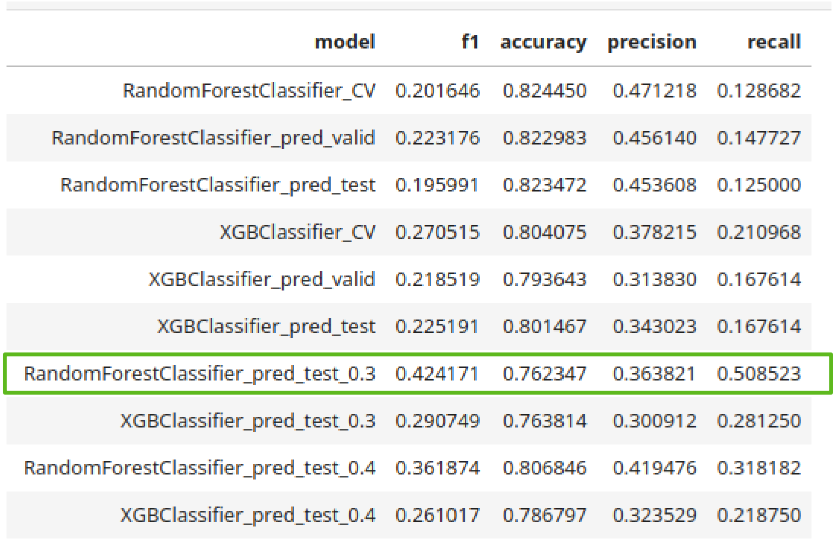
Methodology

All features were used for modeling. New features were derived. The models were developed with sets: 60/20/20 for train/validation/test.

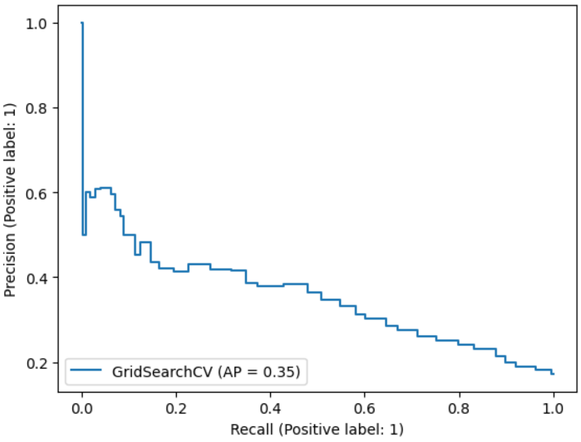
Models observations

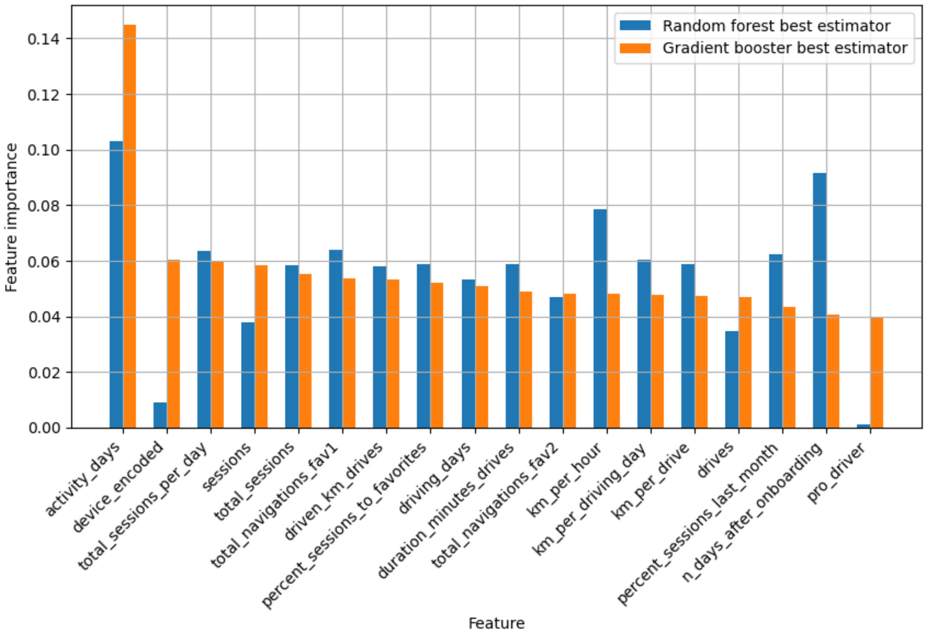
Both models bring similar poor churn-prediction power: e.g 59 matches out of 352 for GBM

- It is good at predicting retention: 1580 out of 1693
- Models are not well explainable, specially GBM.
- Most influential feature is *activity_days for both models*. This was also the case for LG. *Further feature importances are discrepant between models*.
- Further tuning with probability threshold set to 0.4 improves FN effects notably for RF. RF also has a better precision-recall curve than GBM



model	f1	accuracy	precision	recall
RandomForestClassifier_CV	0.201646	0.824450	0.471218	0.128682
RandomForestClassifier_pred_valid	0.223176	0.822983	0.456140	0.147727
RandomForestClassifier_pred_test	0.195991	0.823472	0.453608	0.125000
XGBClassifier_CV	0.270515	0.804075	0.378215	0.210968
XGBClassifier_pred_valid	0.218519	0.793643	0.313830	0.167614
XGBClassifier_pred_test	0.225191	0.801467	0.343023	0.167614
RandomForestClassifier_pred_test_0.3	0.424171	0.762347	0.363821	0.508523
XGBClassifier_pred_test_0.3	0.290749	0.763814	0.300912	0.281250
RandomForestClassifier_pred_test_0.4	0.361874	0.806846	0.419476	0.318182
XGBClassifier_pred_test_0.4	0.261017	0.786797	0.323529	0.218750





Next Steps

- Get more data: long term, new features: geolocation, etc.
- Clarify meaning of some features: session vs. navigation: no clear definition
- Perform PCA for feature selection