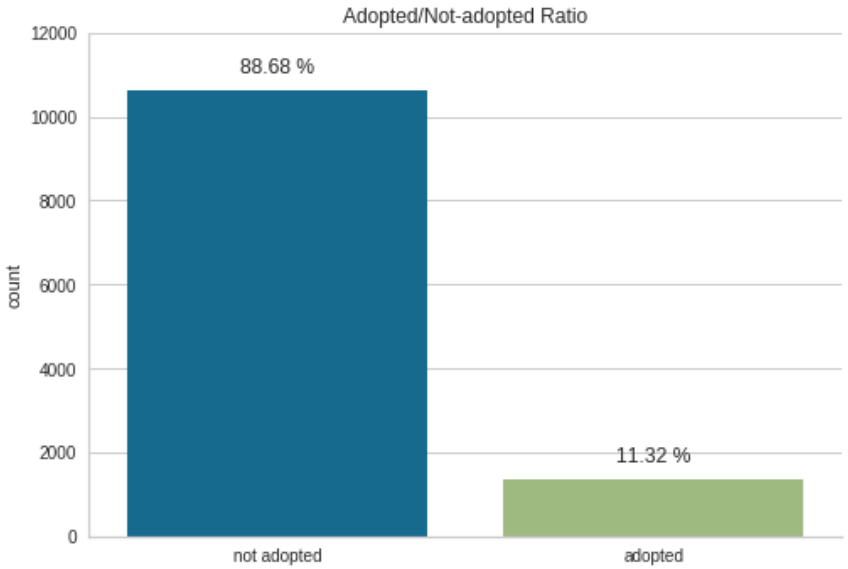
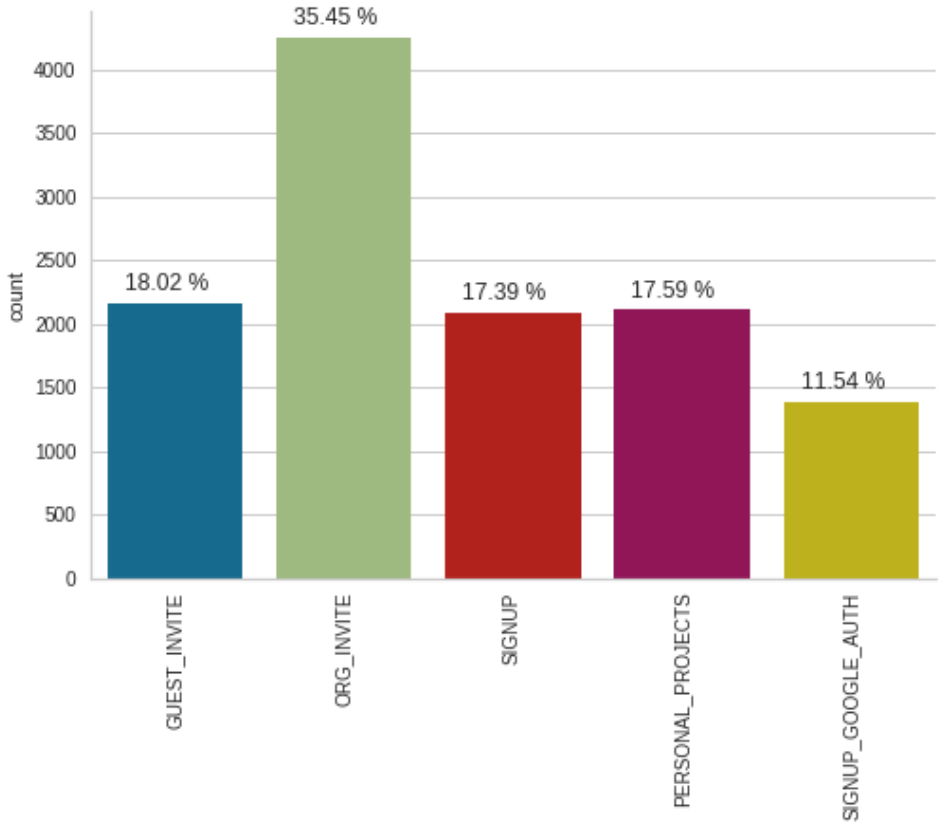
After combining the two csv files, adopted users are identified and labeled. Adopted users are defined as users who have logged into the product on three separate days in at least one seven day period.

This dataset is imbalanced. Only 11.32% of all users adopted.



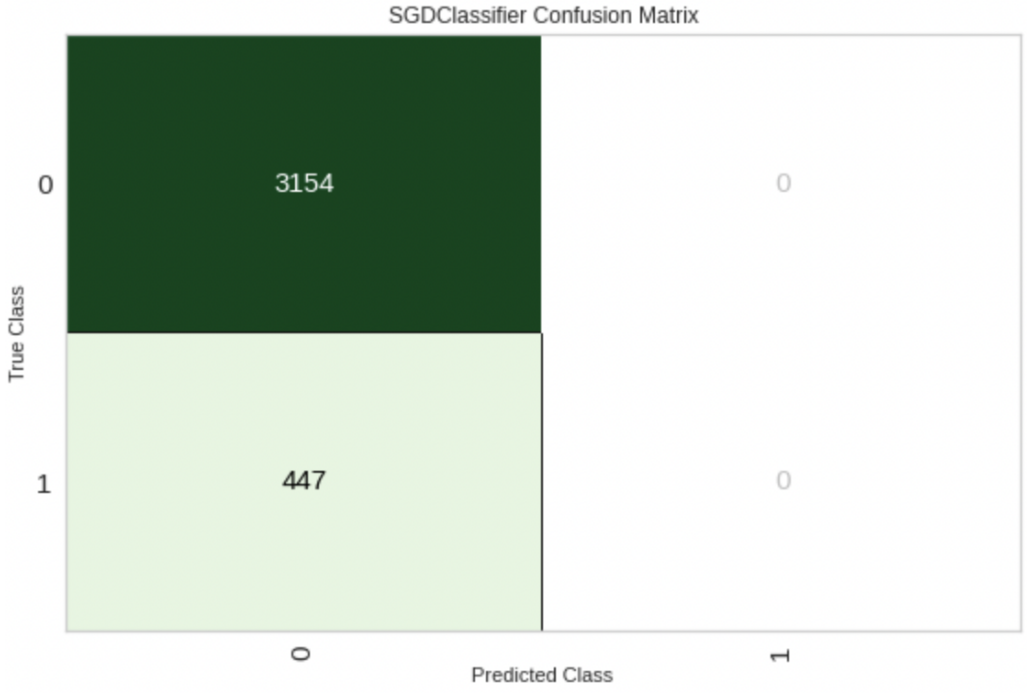
More than one-third (35.45%) of users are invited through their organization.

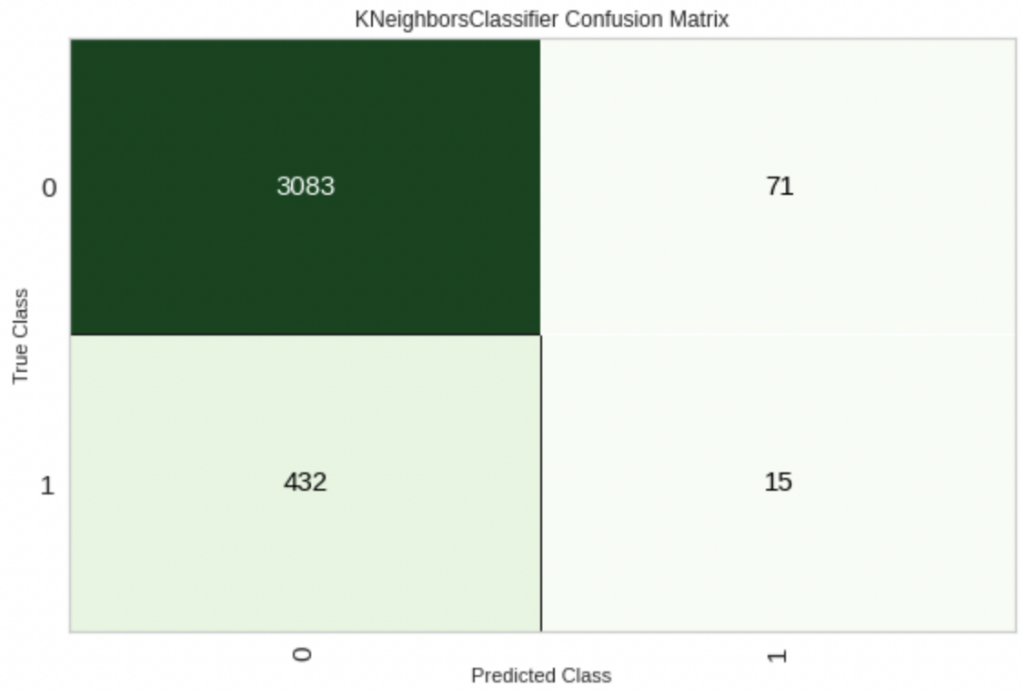


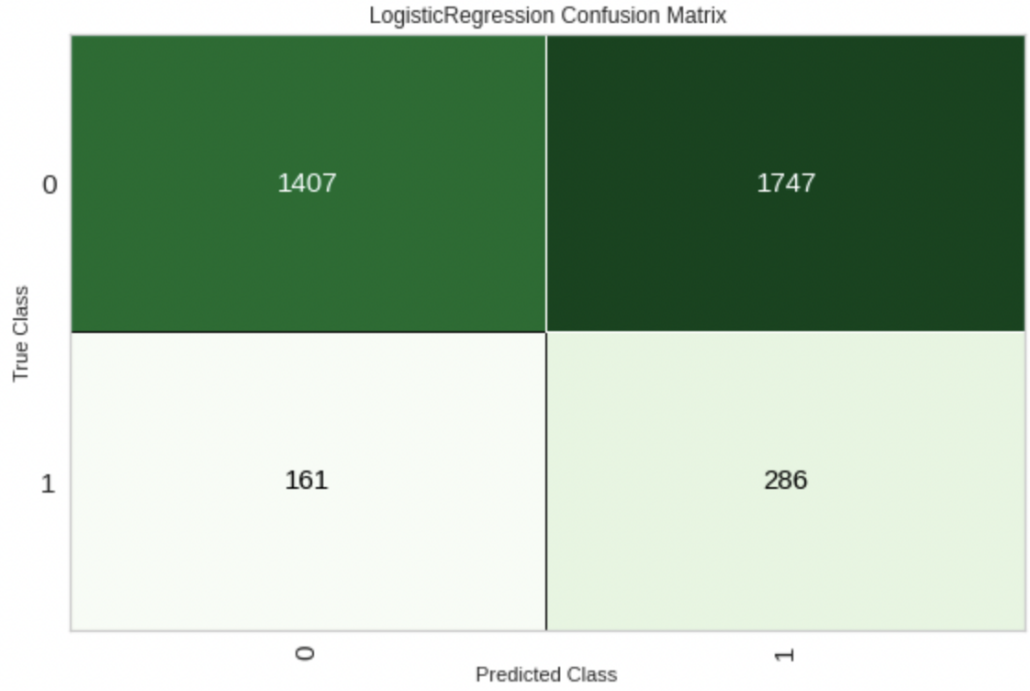
In order to find out which classification would work best for our data, I used pycaret to compare models. Pycaret setup also included SMOTE to fix the imbalance issue.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
knn	K Neighbors Classifier	0.8028	0.5148	0.1328	0.1286	0.1121	0.0167	0.0185	0.278
dt	Decision Tree Classifier	0.5480	0.5548	0.4995	0.1200	0.1931	0.0225	0.0336	0.034
et	Extra Trees Classifier	0.5480	0.5548	0.4995	0.1200	0.1931	0.0225	0.0336	0.629
lightgbm	Light Gradient Boosting Machine	0.5480	0.5548	0.4995	0.1200	0.1931	0.0225	0.0336	0.181
rf	Random Forest Classifier	0.5467	0.5543	0.5006	0.1199	0.1930	0.0222	0.0332	0.746
gbc	Gradient Boosting Classifier	0.5423	0.5564	0.5083	0.1199	0.1935	0.0224	0.0343	0.423
ada	Ada Boost Classifier	0.5061	0.5674	0.5981	0.1259	0.2079	0.0350	0.0580	0.287
lr	Logistic Regression	0.5054	0.5702	0.6003	0.1260	0.2082	0.0354	0.0587	0.312
ridge	Ridge Classifier	0.5054	0.0000	0.6003	0.1260	0.2082	0.0354	0.0587	0.039
lda	Linear Discriminant Analysis	0.5054	0.5702	0.6003	0.1260	0.2082	0.0354	0.0587	0.045
nb	Naive Bayes	0.4010	0.5698	0.7309	0.1226	0.2091	0.0294	0.0624	0.033
qda	Quadratic Discriminant Analysis	0.3915	0.5180	0.6802	0.1135	0.1924	0.0109	0.0254	0.036
svm	SVM - Linear Kernel	0.3736	0.0000	0.7726	0.1232	0.2110	0.0307	0.0712	0.072

Performance is not great across the board. I focused on 3 models: K Neighbors Classifier (best accuracy score), SVM - Linear Kernel (best recall score) and Logistic Regression (good compromise)







Tuned SVM model optimizes accuracy by only guessing all users won't adopt.
The tuned Knn model is prone to false negatives while logistic Regression model is prone to false positives.

There isn't a factor that predicts user adoption. This dataset seems random.