Customer Churn Prediction

By Derek Plemons

Table of Contents

Business Problem

- Can we predict which customers are most likely to churn?

Datasets

1. Telecom Churn Dataset
    a. 7031 Rows
    b. 21 Columns
    c. Telco-Customer-Churn.csv retrieved from kaggle.com

Features

All Features of Telecom Dataset

- Gender, SeniorCitizen, Partner, Dependents, Tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Churn

- We will be predicting on the Churn column

Modeling

Used Pycaret to determine best model

Model of Choice:

ADA Boost Classifier

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ada	Ada Boost Classifier	0.8006	0.8427	0.5309	0.6727	0.5918	0.4624	0.4692	0.0400
lr	Logistic Regression	0.8004	0.8379	0.5391	0.6662	0.5947	0.4644	0.4698	0.4560
ridge	Ridge Classifier	0.7982	0.0000	0.5153	0.6689	0.5813	0.4514	0.4585	0.0050
gbc	Gradient Boosting Classifier	0.7974	0.8417	0.5257	0.6624	0.5853	0.4537	0.4596	0.0990
lda	Linear Discriminant Analysis	0.7937	0.8305	0.5489	0.6425	0.5911	0.4544	0.4576	0.0100
catboost	CatBoost Classifier	0.7917	0.8328	0.5242	0.6451	0.5780	0.4417	0.4462	1.2370
lightgbm	Light Gradient Boosting Machine	0.7882	0.8242	0.5279	0.6337	0.5753	0.4359	0.4396	0.1910
rf	Random Forest Classifier	0.7819	0.8178	0.4952	0.6260	0.5525	0.4110	0.4162	0.0980
xgboost	Extreme Gradient Boosting	0.7728	0.8114	0.5086	0.5974	0.5485	0.3983	0.4011	0.3440
et	Extra Trees Classifier	0.7716	0.7895	0.4854	0.6002	0.5361	0.3869	0.3912	0.1050
svm	SVM - Linear Kernel	0.7669	0.0000	0.3857	0.6289	0.4519	0.3243	0.3482	0.0130
knn	K Neighbors Classifier	0.7590	0.7413	0.4526	0.5746	0.5058	0.3494	0.3541	0.2040
nb	Naive Bayes	0.7396	0.8235	0.7599	0.5150	0.6136	0.4281	0.4468	0.0050
dt	Decision Tree Classifier	0.7114	0.6418	0.4884	0.4707	0.4792	0.2797	0.2799	0.0070
qda	Quadratic Discriminant Analysis	0.6095	0.6401	0.7071	0.3959	0.4892	0.2239	0.2673	0.0060

ADA Boost Classifier

Accuracy: 0.8006

AUC: 0.8427

Recall: 0.5309

Precision: 0.6727

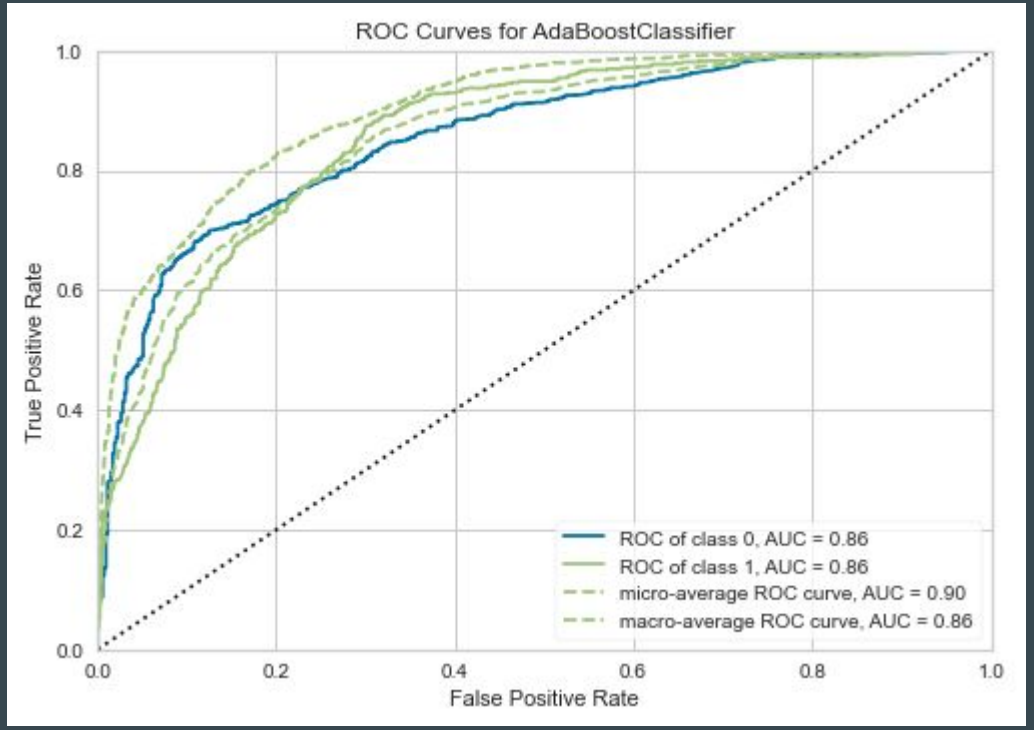
F1: 0.5918

Kappa: 0.4624

MCC: 0.4692

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7992	0.8339	0.4851	0.6842	0.5677	0.4418	0.4529
1	0.8154	0.8649	0.5075	0.7312	0.5991	0.4843	0.4979
2	0.7911	0.8369	0.4925	0.6535	0.5617	0.4281	0.4355
3	0.7627	0.8152	0.5075	0.5714	0.5375	0.3787	0.3799
4	0.8174	0.8496	0.5149	0.7340	0.6053	0.4912	0.5043
5	0.8276	0.8801	0.5373	0.7579	0.6288	0.5207	0.5338
6	0.8195	0.8432	0.6194	0.6860	0.6510	0.5297	0.5309
7	0.8032	0.8460	0.5672	0.6609	0.6104	0.4799	0.4824
8	0.7708	0.8156	0.5299	0.5868	0.5569	0.4028	0.4038
9	0.7992	0.8418	0.5481	0.6607	0.5992	0.4668	0.4704
Mean	0.8006	0.8427	0.5309	0.6727	0.5918	0.4624	0.4692
SD	0.0200	0.0188	0.0380	0.0575	0.0334	0.0466	0.0489

ROC Curves for ADA Boost Classifier



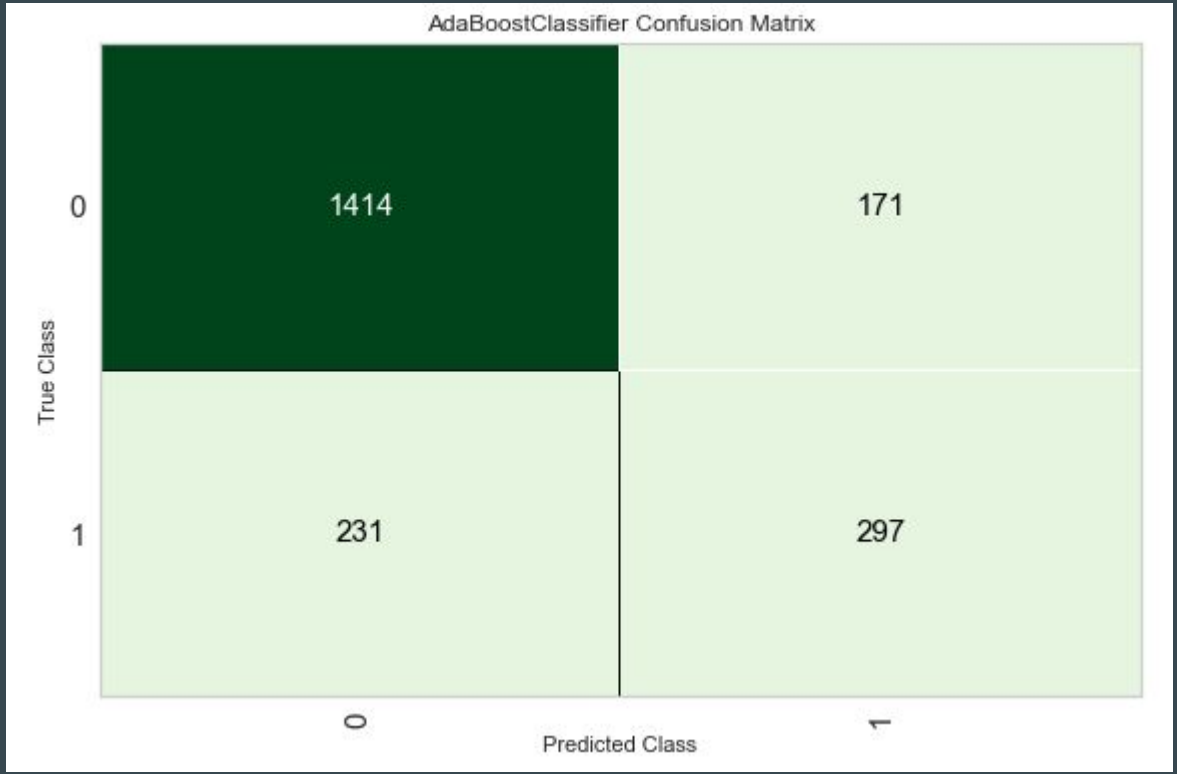
ADA Boost Classifier Confusion Matrix

True Positives: 1414

True Negatives: 297

False Positives: 231

False Negatives: 171



Conclusion

Using the Telco-Customer-churn dataset from kaggle, we can predict which customers are likely to churn with an 80% accuracy with the given features.