



# Telecom Churn Classification Analysis

Brennan Mathis

## Problem Statement

Use customer usage data, location-based data, and plan information to assess affect on customer retention

## Methodology

- XGBoost
- Logistic Regression
- Stochastic Gradient Descent Classification
- KNN
- Random Forest
- Decision Tree with Bagging Classifier
- AdaBoost
- SVM/SVC

RandomForestClassifier(bootstrap=True, class\_weight='balanced', criterion='entropy', max\_depth=None, max\_features=30, max\_leaf\_nodes=None, min\_impurity\_decrease=0.0, min\_impurity\_split=None, min\_samples\_leaf=1, min\_samples\_split=2, min\_weight\_fraction\_leaf=0.0, n\_estimators=10, n\_jobs=-1, oob\_score=False, random\_state=None, verbose=0, warm\_start=False) precision recall f1-score support False 0.99 1.00 1.00 2280 True 1.00 0.96 0.98 386 accuracy 0.99 2666 macro avg 1.00 0.98 0.99 2666 weighted avg 0.99 0.99 0.99 2666

KNeighborsClassifier(algorithm='ball\_tree', leaf\_size=30, metric='minkowski', metric\_params=None, n\_jobs=None, n\_neighbors=5, p=2, weights='uniform') precision recall f1-score support False 0.89 0.99 0.94 2280 True 0.82 0.25 0.38 386 accuracy 0.88 2666 macro avg 0.85 0.62 0.66 2666 weighted avg 0.88 0.88 0.86 2666

LogisticRegression(C=1.0, class\_weight=None, dual=False, fit\_intercept=True, intercept\_scaling=1, l1\_ratio=None, max\_iter=1000, multi\_class='warn', n\_jobs=None, penalty='l2', random\_state=42, solver='newton-cg', tol=0.0001, verbose=0, warm\_start=False) precision recall f1-score support False 0.88 0.98 0.93 2280 True 0.64 0.22 0.33 386

accuracy 0.87 2666 macro avg 0.76 0.60 0.63 2666 weighted avg 0.85 0.87 0.84 2666

SGDClassifier(alpha=0.0001, average=False, class\_weight=None, early\_stopping=False, epsilon=0.1, eta0=0.0, fit\_intercept=True, l1\_ratio=0.15, learning\_rate='optimal', loss='hinge', max\_iter=1000, n\_iter\_no\_change=5, n\_jobs=None, penalty='l2', power\_t=0.5, random\_state=None, shuffle=True, tol=0.001, validation\_fraction=0.1, verbose=0, warm\_start=False) precision recall f1-score support False 0.87 0.99 0.92 2280 True 0.65 0.09 0.16 386 accuracy 0.86 2666 macro avg 0.76 0.54 0.54 2666 weighted avg 0.83 0.86 0.81 2666

DecisionTreeClassifier(class\_weight=None, criterion='entropy', max\_depth=10, max\_features=None, max\_leaf\_nodes=None, min\_impurity\_decrease=0.0, min\_impurity\_split=None, min\_samples\_leaf=1, min\_samples\_split=20, min\_weight\_fraction\_leaf=0.0, presort=False, random\_state=None, splitter='best') precision recall f1-score support False 0.97 0.99 0.98 2280 True 0.94 0.82 0.88 386 accuracy 0.97 2666 macro avg 0.95 0.91 0.93 2666 weighted avg 0.97 0.97 0.97 2666

SVC(C=1000, cache\_size=200, class\_weight=None, coef0=0.0, decision\_function\_shape='ovr', degree=1, gamma='auto\_deprecated', kernel='rbf', max\_iter=-1, probability=False, random\_state=None, shrinking=True, tol=0.001, verbose=False) precision recall f1-score support False 0.93 0.99 0.96 2280 True 0.93 0.53 0.68 386 accuracy 0.93 2666 macro avg 0.93 0.76 0.82 2666 weighted avg 0.93 0.93 0.92 2666

XGBClassifier(alpha=0.01, base\_score=0.5, booster='gbtree', colsample\_bylevel=1, colsample\_bynode=1, colsample\_bytree=1, gamma=0, learning\_rate=0.1, max\_delta\_step=0, max\_depth=3, min\_child\_weight=1, missing=None, n\_estimators=100, n\_jobs=1, nthread=None, objective='binary:logistic', random\_state=0, reg\_alpha=0, reg\_lambda=1, scale\_pos\_weight=1, seed=None, silent=None, subsample=1, verbosity=1) precision recall f1-score support False 0.97 1.00 0.98 2280 True 0.97 0.80 0.87 386 accuracy 0.97 2666 macro avg 0.97 0.90 0.93 2666 weighted avg 0.97 0.97 0.96 2666

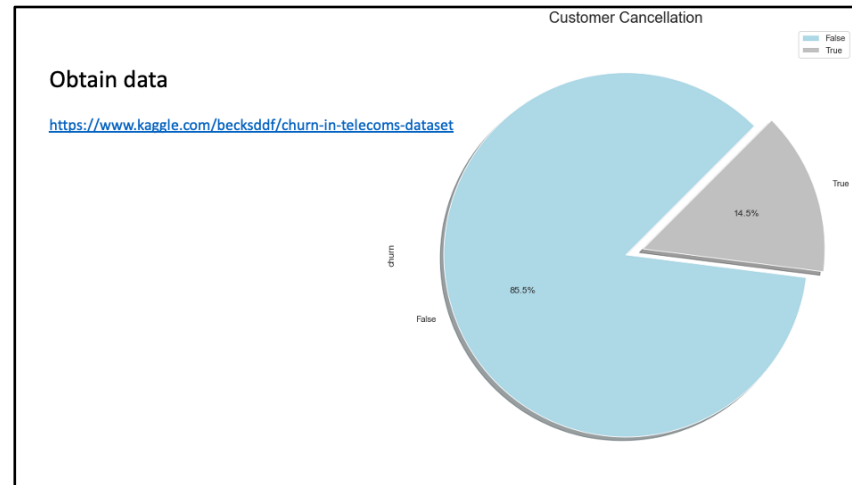
AdaBoostClassifier(algorithm='SAMME.R', base\_estimator=None, learning\_rate=1, n\_estimators=50, random\_state=None) precision recall f1-score support False 0.92 0.97 0.94 2280 True 0.71 0.47 0.57 386 accuracy 0.90 2666 macro avg 0.81 0.72 0.76 2666 weighted avg 0.89 0.90 0.89 2666

## Business Value

- To predict customer's potential for cancelling service, so as to avoid losing customers
- Retaining customer is less costly than obtaining new customers, due to advertising costs, initial contract 'deal' costs, etc

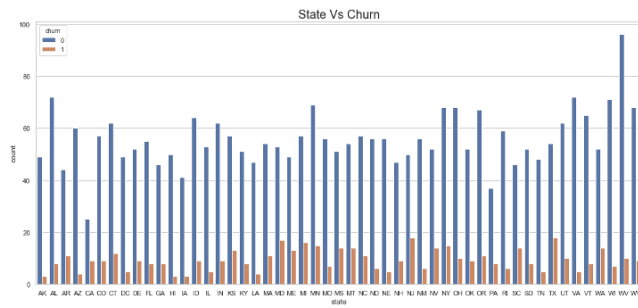
Based on dataset with the following info:

'state', 'account length', 'area code', 'phone number',  
'international plan', 'voice mail plan', 'number vmail messages',  
'total day minutes', 'total day calls', 'total day charge',  
'total eve minutes', 'total eve calls', 'total eve charge',  
'total night minutes', 'total night calls', 'total night charge',  
'total intl minutes', 'total intl calls', 'total intl charge',  
'customer service calls', 'churn'



Columns = ['state', 'account length', 'area code', 'phone number', 'international plan', 'voice mail plan', 'number vmail messages', 'total day minutes', 'total day calls', 'total day charge', 'total eve minutes', 'total eve calls', 'total eve charge', 'total night minutes', 'total night calls', 'total night charge', 'total intl minutes', 'total intl calls', 'total intl charge', 'customer service calls', 'churn'],

## Location Assessment



## Factors Assessed

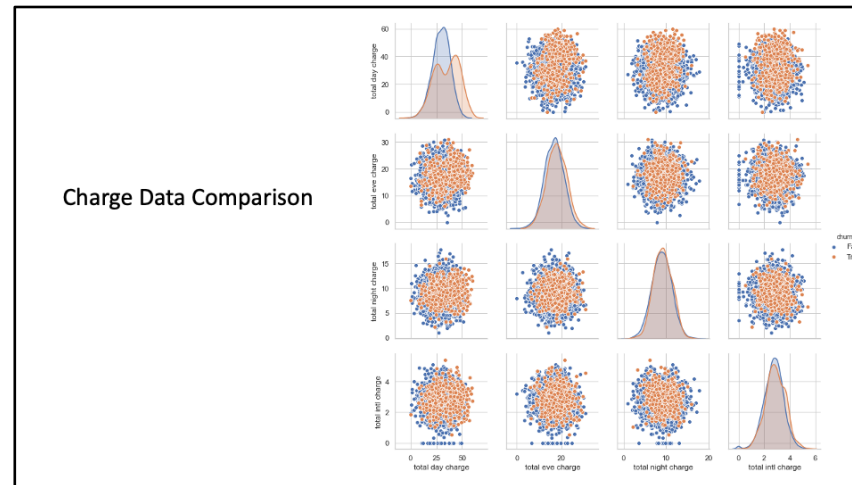
- 'state', 'account length', 'area code', 'phone number',
- 'international plan', 'voice mail plan', 'number vmail messages',
- 'total day minutes', 'total day calls', 'total day charge',
- 'total eve minutes', 'total eve calls', 'total eve charge',
- 'total night minutes', 'total night calls', 'total night charge',
- 'total intl minutes', 'total intl calls', 'total intl charge',
- 'customer service calls', 'churn'

'state', 'account length', 'area code', 'phone number',  
'international plan', 'voice mail plan', 'number vmail messages',  
'total day minutes', 'total day calls', 'total day charge',  
'total eve minutes', 'total eve calls', 'total eve charge',  
'total night minutes', 'total night calls', 'total night charge',  
'total intl minutes', 'total intl calls', 'total intl charge',  
'customer service calls', 'churn'



[illegible]

```
'phone number'  
'area code',  
'number vmail messages',  
'total day minutes',  
'total eve minutes',  
'total night minutes',  
'total intl minutes',
```

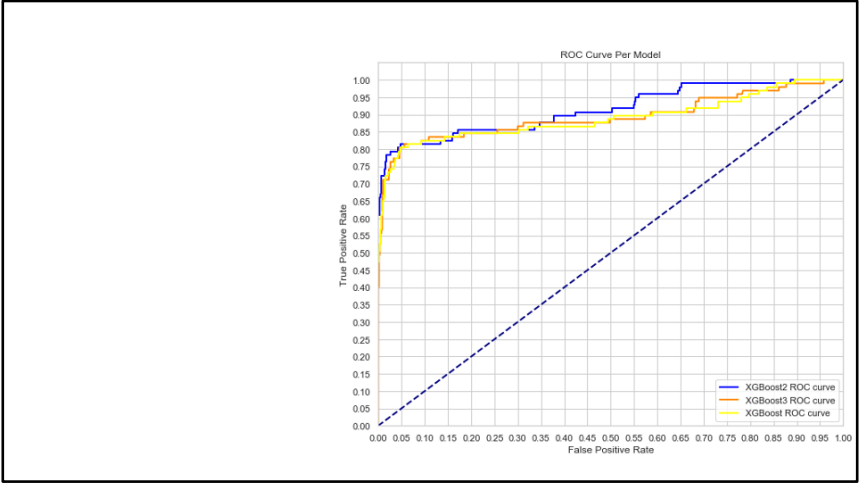


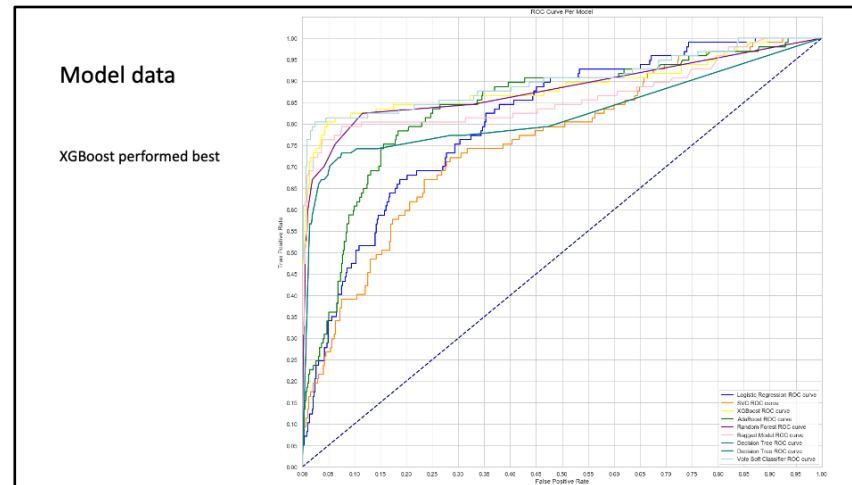
Total day charge, total evening charge, total international charge are factors that affect churn rate  
Total day charge has most effect



## Business Recommendations

- Providing positive customer service call experience
- Providing positive customer service billing inquiries experience
- Assess plans provided to new customers to ensure best plan for their use is offered to them
- Assess reliability of towers and service in states with higher churn rate





Test AUC:

XGBoost: 0.914414903237475

Logistic Regression: 0.8083740278531379

SVC: 0.7582202930005426

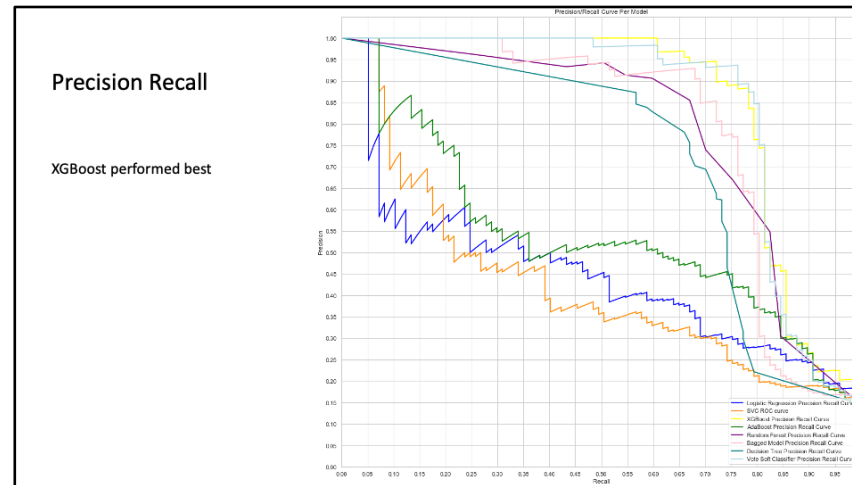
AdaBoost: 0.8404503526858382

Random Forest: 0.8795351781515645

Bagged Model: 0.8618556701030927

Decision Tree: 0.8221920781334779

Voting Classifier: 0.9007958039428468



Average Precision/Recall Score:

XGBoost: 0.85

Logistic Regression: 0.45

SVC: 0.42

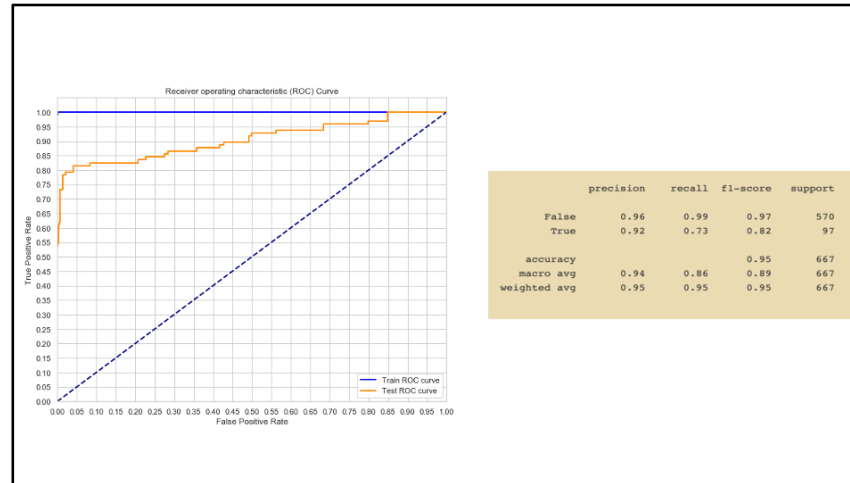
AdaBoost: 0.54

Random Forest: 0.75

Bagged Model: 0.79

Decision Tree: 0.67

Voting Classifier: 0.84



Voting classifier

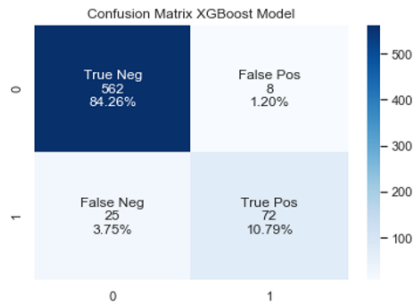
Train AUC: 0.9993977820198164

Test AUC: 0.914414903237475

Matthew's Correlation Coefficient 0.7832481210271933

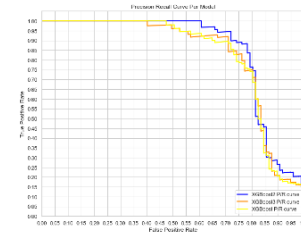
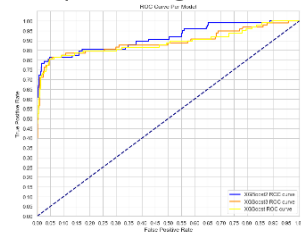


## Interpret data



## Future Analysis

- Assess and update hyperparameters to reduce False Negatives
- Test on alternate similar datasets
- Explore voting classifier combining algorithms



Upon exploring XGBoost, changing parameters changed overall results in both accuracy and precision/recall score.

