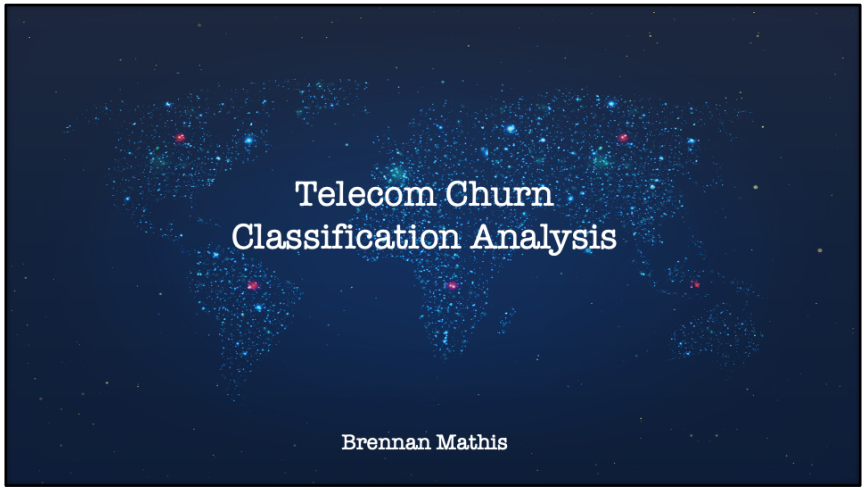

Telecom Churn
Classification Analysis
Brennan Mathis

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

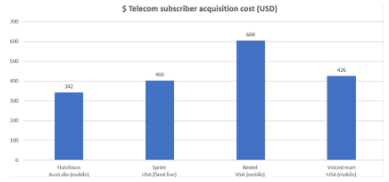
SVC(C=1000, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=1, gamma='auto_deprecated', kernel='rbf', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False) precision recall f1-score support False 0.93 0.99 0.96 2280 True 0.93 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

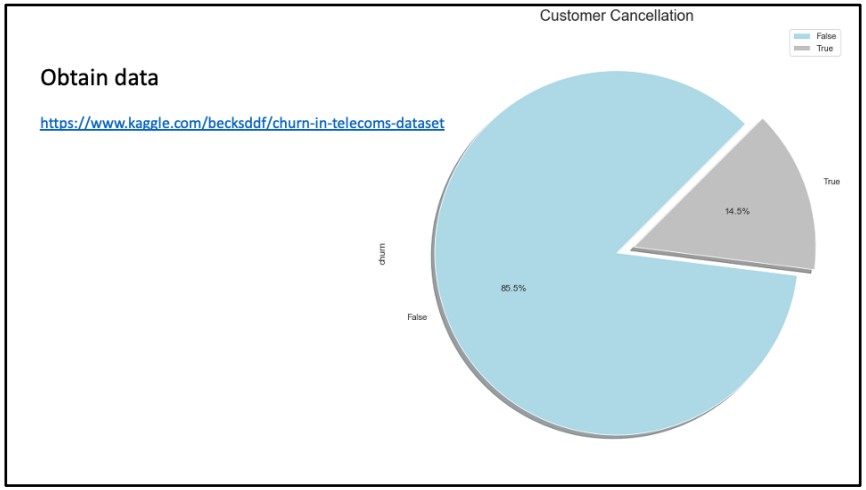
Business Value

- 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, and overall customer acquisition costs (CAC)



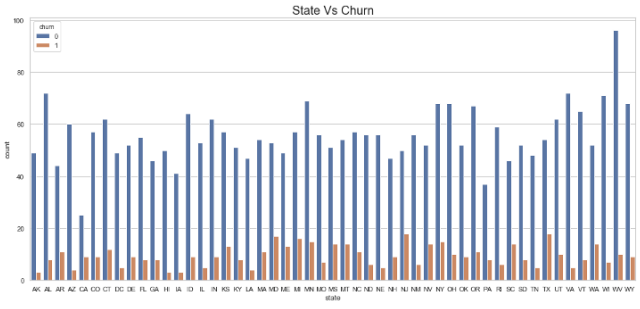
Based on dataset with the following info:
'state', 'account length', 'area code', 'phone number',
'international plan', 'voice mail plan', 'number vmail messages',
'total day minutes', 'total day calls', 'total day charge',
'total eve minutes', 'total eve calls', 'total eve charge',
'total night minutes', 'total night calls', 'total night charge',
'total intl minutes', 'total intl calls', 'total intl charge',
'customer service calls', 'churn'

**image https://www.performancemagazine.org/june-smartkpi-telecom-sub-acquisition/



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'],

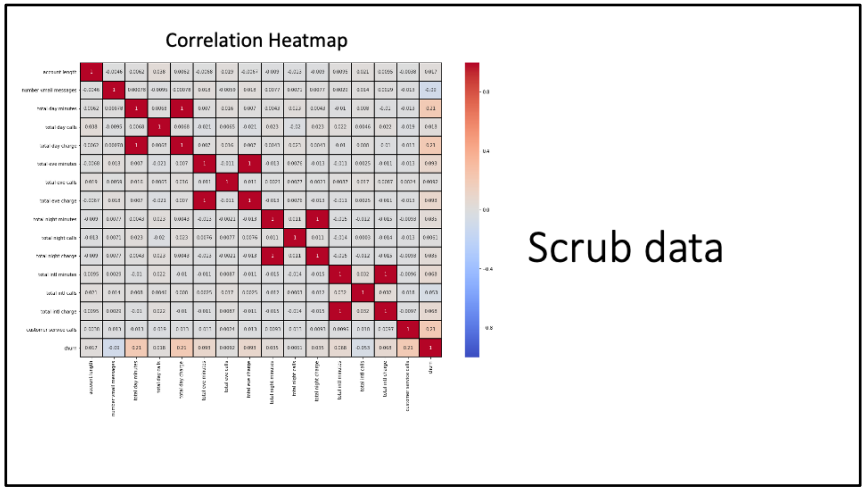
Location Assessment



Factors Assessed

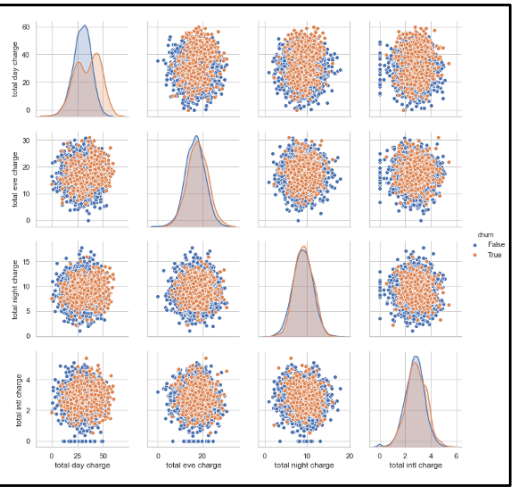
- 'state', 'account length', 'area code', 'phone number',
- 'international plan', 'voice mail plan', 'number vmail messages',
- 'total day minutes', 'total day calls', 'total day charge',
- 'total eve minutes', 'total eve calls', 'total eve charge',
- 'total night minutes', 'total night calls', 'total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

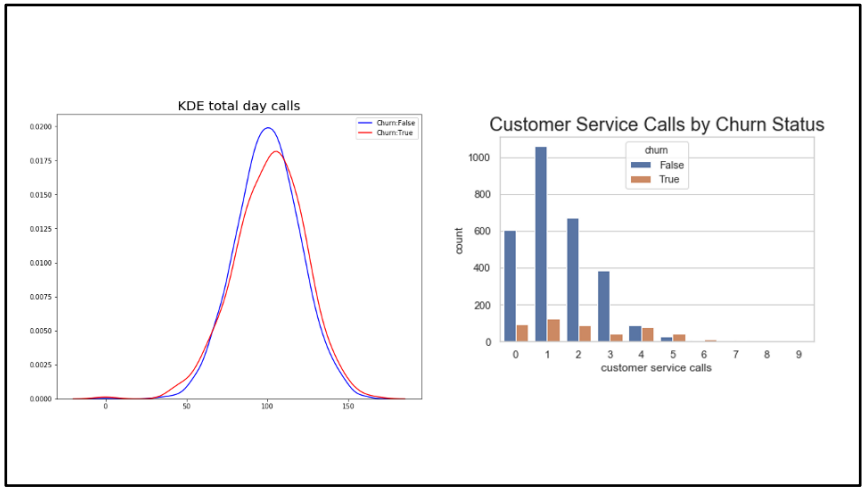


Removed following data due to correlation or relevancy issues
'phone number'
'area code',
'number vmail messages',
'total day minutes',
'total eve minutes',
'total night minutes',
'total intl minutes',

Charge Data Comparison



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


KDE total day calls
Churn:False
Churn:True
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
50
100
150
Customer Service Calls by Churn Status
churn
False
True
count
1000
800
600
400
200
0
0
1
2
3
4
5
6
7
8
9
customer service calls

Business Recommendations

PROVIDING POSITIVE CUSTOMER SERVICE CALL EXPERIENCE

PROVIDING POSITIVE CUSTOMER SERVICE BILLING INQUIRIES EXPERIENCE

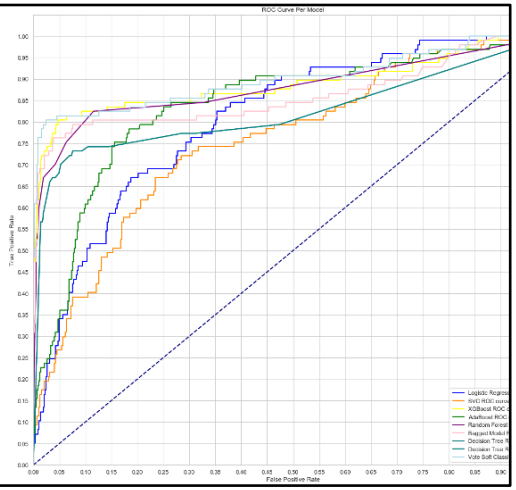
ASSESS PLANS PROVIDED TO NEW CUSTOMERS TO ENSURE BEST PLAN FOR THEIR USE IS OFFERED TO THEM

ASSESS RELIABILITY OF TOWERS AND SERVICE IN STATES WITH HIGHER CHURN RATE

Model data

ROC Curve

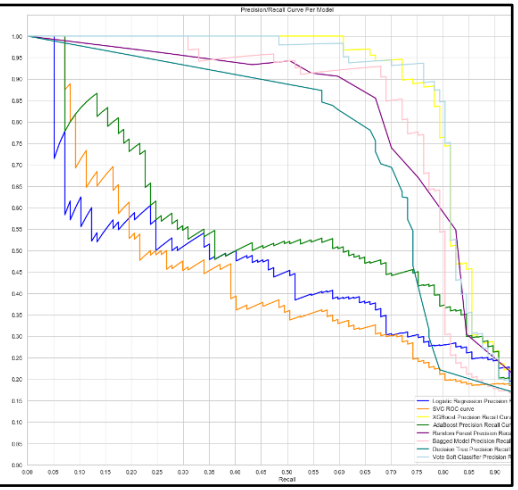
XGBoost performed best



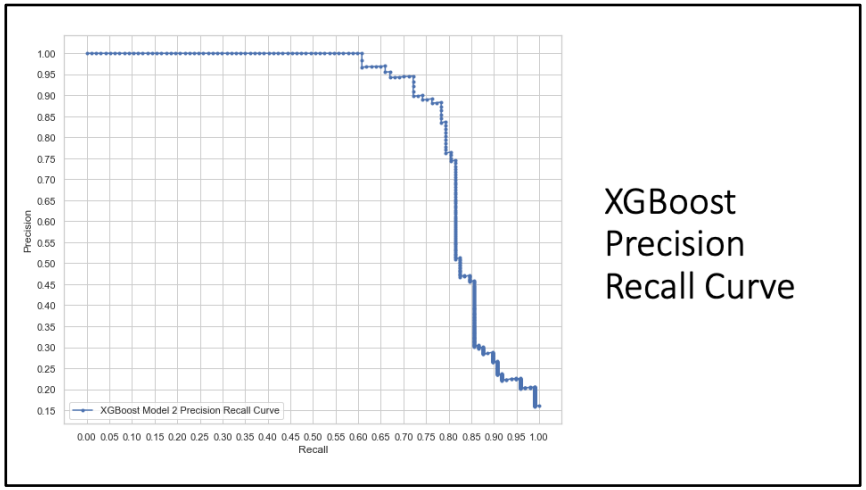
Test AUC:
XGBoost: 0.914414903237475
Logistic Regression: 0.8083740278531379
SVC: 0.7582202930005426
AdaBoost: 0.8404503526858382
Random Forest: 0.8795351781515645
Bagged Model:0.8618556701030927
Decision Tree: 0.8221920781334779
Voting Classifier: 0.9007958039428468

Precision Recall

XGBoost performed best



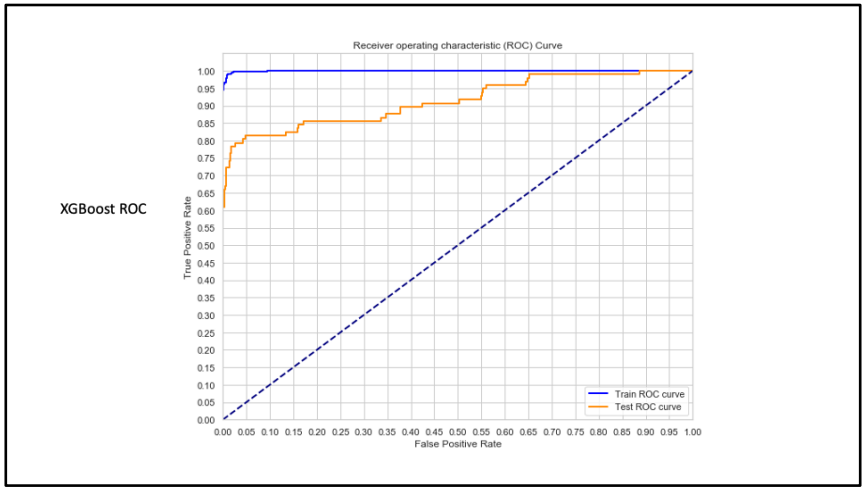
Average Precision/Recall Score:
XGBoost: 0.85
Logistic Regression: 0.45
SVC: 0.42
AdaBoost: 0.54
Random Forest: 0.75
Bagged Model:0.79
Decision Tree: 0.67
Voting Classifier: 0.84



Average precision-recall score: 0.85

Prefer this method due to fact that there are unequal number of observations/ moderate imbalance

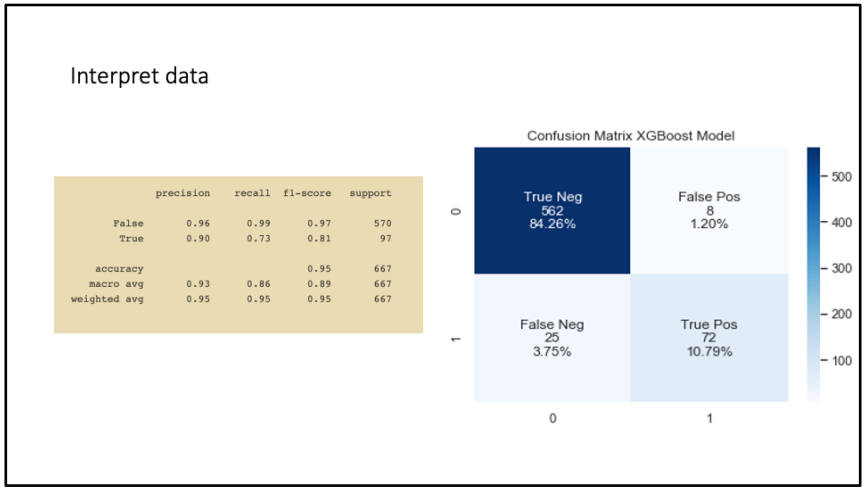
XGBoost ROC



Train AUC: 0.9993977820198164 Test AUC: 0.914414903237475
True pos = sensitivity

Interpret data

	precision	recall	f1-score	support
False	0.96	0.99	0.97	570
True	0.90	0.73	0.81	97
accuracy			0.95	667
macro avg	0.93	0.86	0.89	667
weighted avg	0.95	0.95	0.95	667



array([[562, 8],
[25, 72]])

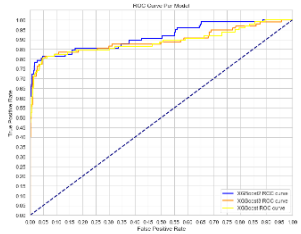
Need to reduce false negatives
False negatives provide harmful info, due to ability to predict cancellations being impossible with this status. At least false positives allude to customer dissatisfaction which it cannot hurt to try to help customers predicted as churn when they are not in fact in that category. It can even be considered preventative. If false negative, then there is no way to correct or provide positive experience ahead of churn

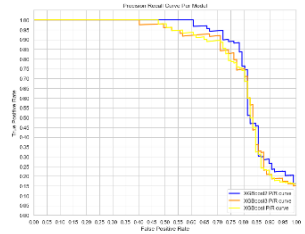
True pos = sensitivity
False pos = false alarm
Recal
Precision is a good measure to determine, when the costs of False Positive is highl shall be the model metric we use to select our best model when there is a high cost associated with False Ne

F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall AND there is an uneven class distribution (large number of Actual Negatives).gative.

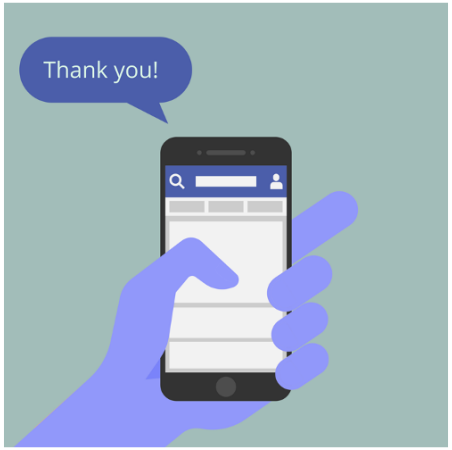
Future Analysis

- 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.


Thank you!