Predicting Telcom Customer Retention Using Machine Learning

Bruno Xie

CHICAGO BOOTH

Introduction

Founded in 1972, with over 40 years of experience in the design and development of high-performance network communications solutions

- Covers network segments: transport, connectivity, and distributed network applications
- Provides services: phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies

Our Goal

Retention management is extremely important as the costs of acquiring new customers are higher than keeping the existing customers

Our goal is help Telco retain customers by predicting users' behaviors

We will run several models including logistic regression, random forest, and boosting to study which factors will impact customer churn

Data Overview

- **Source:** Kaggle
- **Summary:** **7043** Telco customers
  - **Customer Behavior:** churn
  - **Services:** phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
  - **Customer Account Information:** how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
  - **Demographic Info:** gender, age range, and if they have partners and dependents

Data Exploration

Which type of customers are more likely to churn?

Services			
Internet Service	DSL	**Fiber optic**	No
Phone Service	**Yes**		No

Services			
Multiple Lines	Yes	**No**	No phone service
Online Security	Yes	**No**	No internet service
Online Backup	Yes	**No**	No internet service
Device Protection	Yes	**No**	No internet service
Tech Support	Yes	**No**	No internet service
Streaming TV	Yes	**No**	No internet service
Streaming Movie	Yes	**No**	No internet service

CHICAGO BOOTH

TELCO SYSTEMS

Data Exploration

Which type of customers are more likely to churn?

Customer Account Information				
Contract	**Month-to-month**		One year	Two years
Paperless Billing	**Yes**		No	
Payment Method	**Eletronic check**	Mailed check	Bank transfer	Credit card
Monthly Charges	**High**		Low	
Total Charges	High		**Low**	
Tenure	Long		**Short**	

Demographic Info		
Gender	**Female**	Male
Senior Citizen	**Yes**	No
Partner	Yes	**No**
Dependents	Yes	**No**

CHICAGO BOOTH

TELCO SYSTEMS

Data Cleaning

- Dropped irrelevant variable *customer ID*
- Standardized continuous variables *monthly charges, total charges, tenure*
- Transformed categorical variables into separated variables with dummy values
- Split 50% data into training set, 25% into validation set, 25% into test set

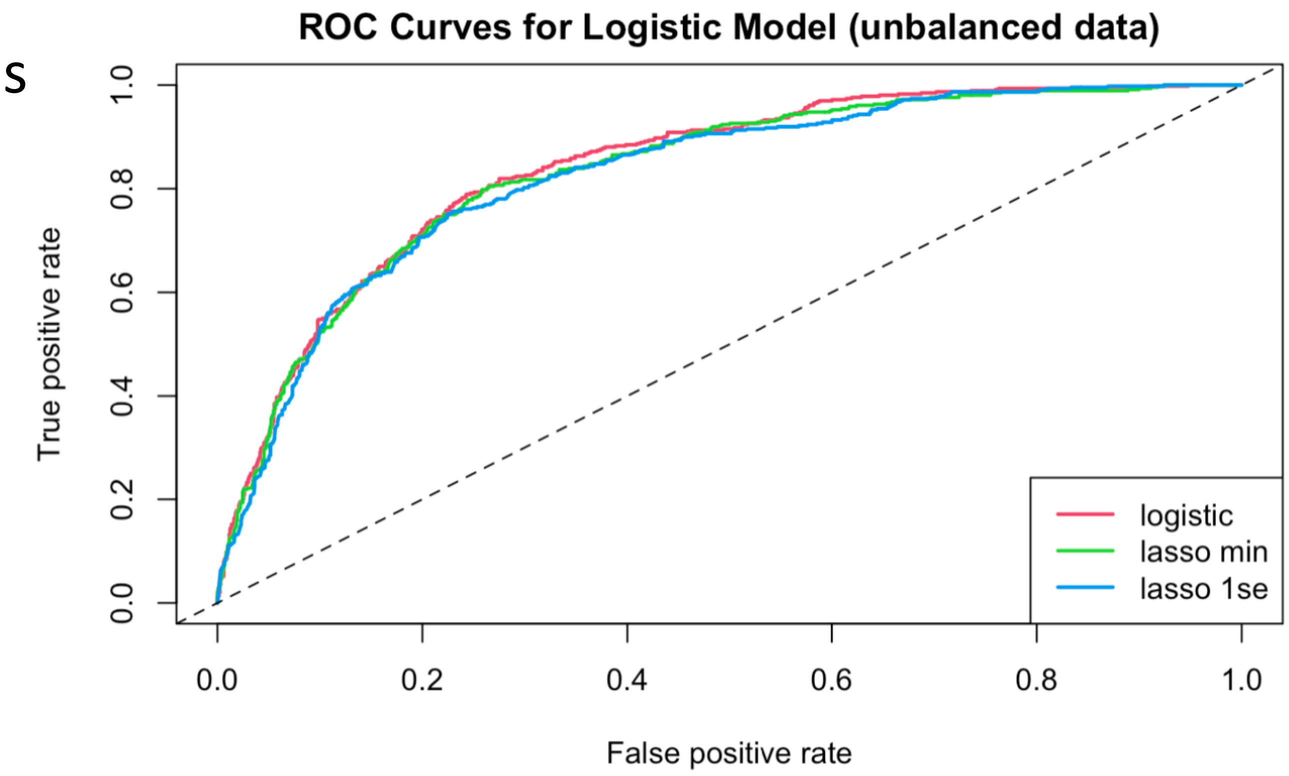
Summary of Results

	Original Data (Unbalanced)			Up-sampling Data (balanced)		
	Deviance	Accuracy (Confusion Matrix)	AUC	Deviance	Accuracy (Confusion Matrix)	AUC
I. Model selection & tuning parameters (based on the validation set)						
Logistic Regression	1482.23	0.8066	0.8399	1757.64	0.7457	0.8380
Lasso min lambda	1484.34	0.7986	0.8318	1755.59	0.7486	0.8383
Lasso 1sd lambda	1489.19	0.7878	0.8257	1758.97	0.7474	0.8348
Random Forest	1503.71	0.8009	0.8334	1606.40	0.7867	0.8238
Boosted Tree	2540.59	0.6752	0.5109	1681.13	0.7537	0.8309
II. Results comparison (based on the test set)						
Logistic Regression	1447.27	0.7981	0.8489	1689.83	0.7497	0.8487
Random Forest	1446.06	0.7986	0.8550	1495.32	0.7901	0.8486
Boosted Tree	1490.41	0.7838	0.8493	1574.47	0.7662	0.8528

Logistic Regression – Unbalanced Data

- Regular logistic regression
- Lasso model with λ that minimizes cross-validated error
- Lasso model with λ in which the error is within 1 se of the minimum

➢ similar predictive powers in terms of deviance, accuracy, and AUC



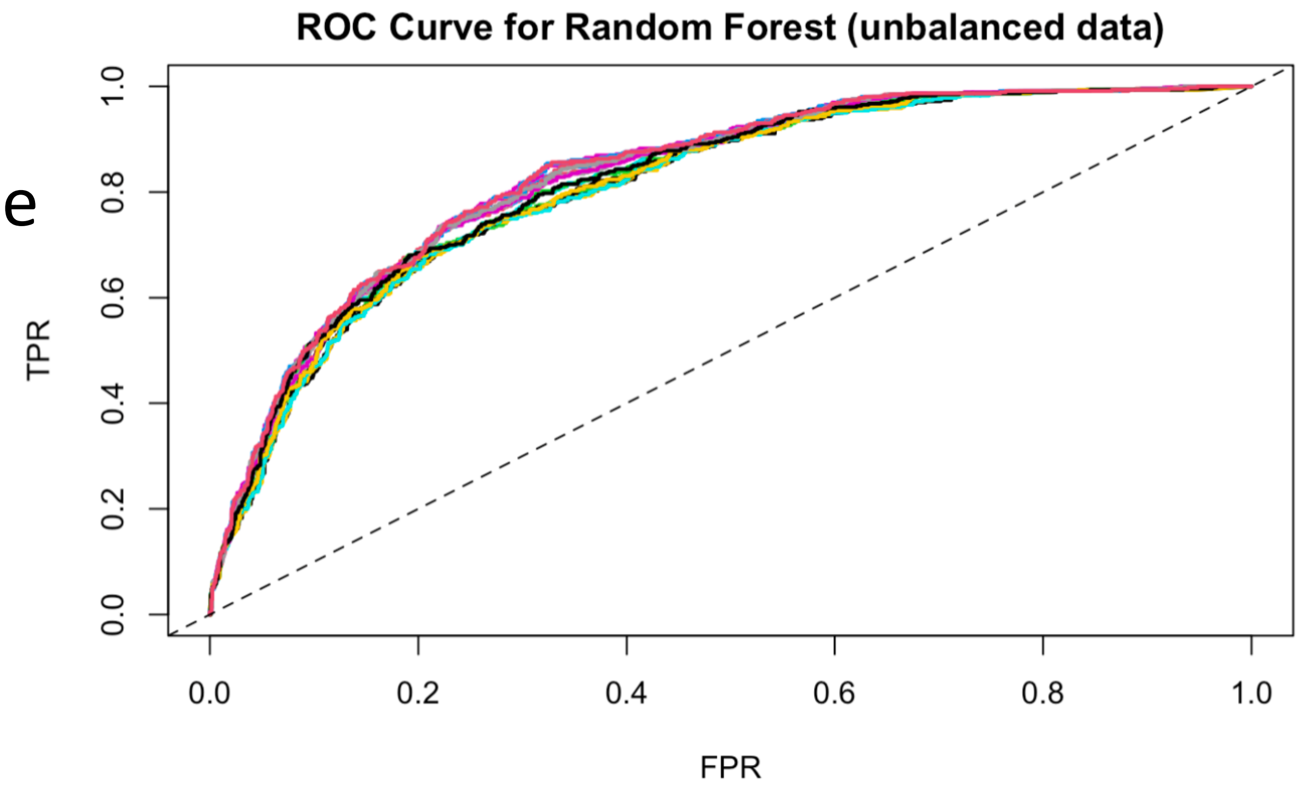
Random Forest – Unbalanced Data

Parameter search:

- number of trees
- number of variables to possibly split at in each node
- minimal node size to tune the parameters

Optimal parameters:

- ➢ number of trees = 1000
- ➢ number of variables to split = 5
- ➢ minimal node size = 5



Boosting – Unbalanced Data

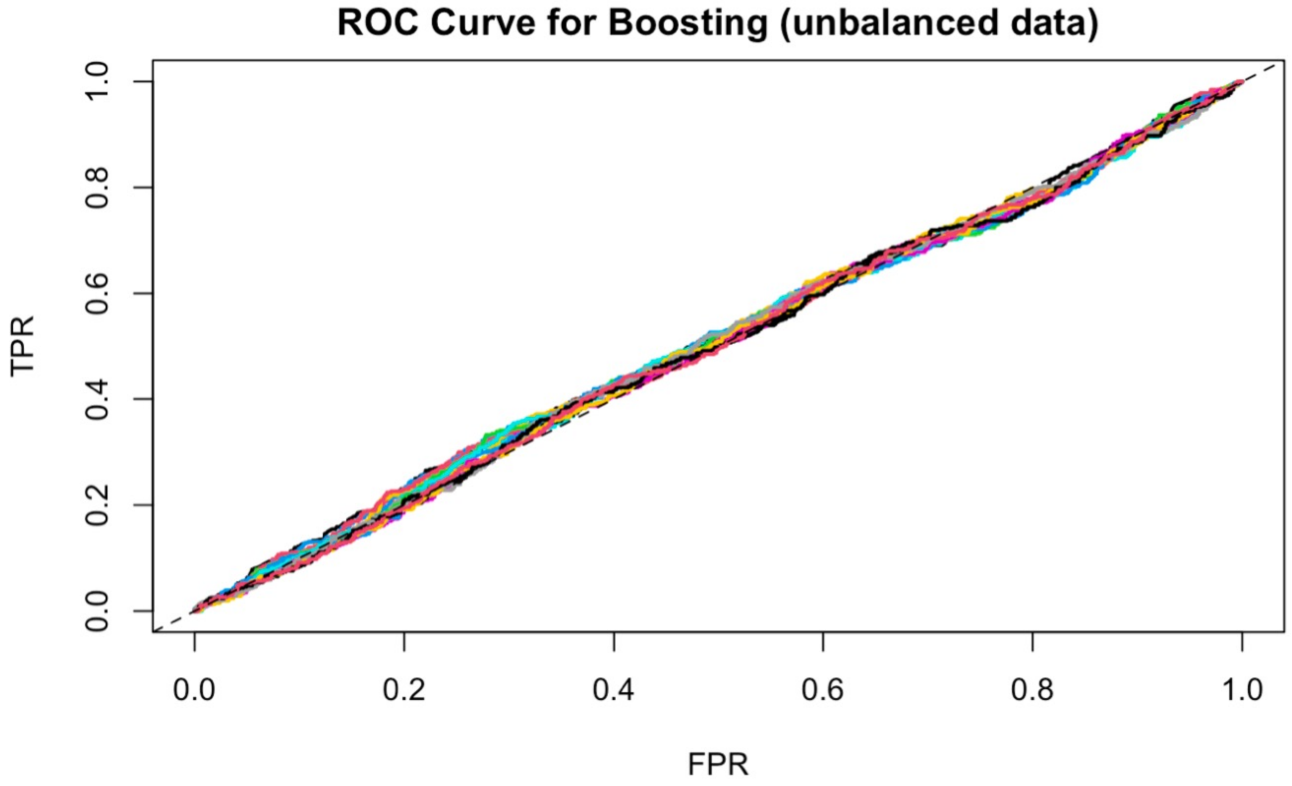
Parameter search:

- learning rate
- maximum depth of trees

Optimal parameters:

➢ learning rate = 0.01

➢ maximum depth of trees = 1



CHICAGO BOOTH

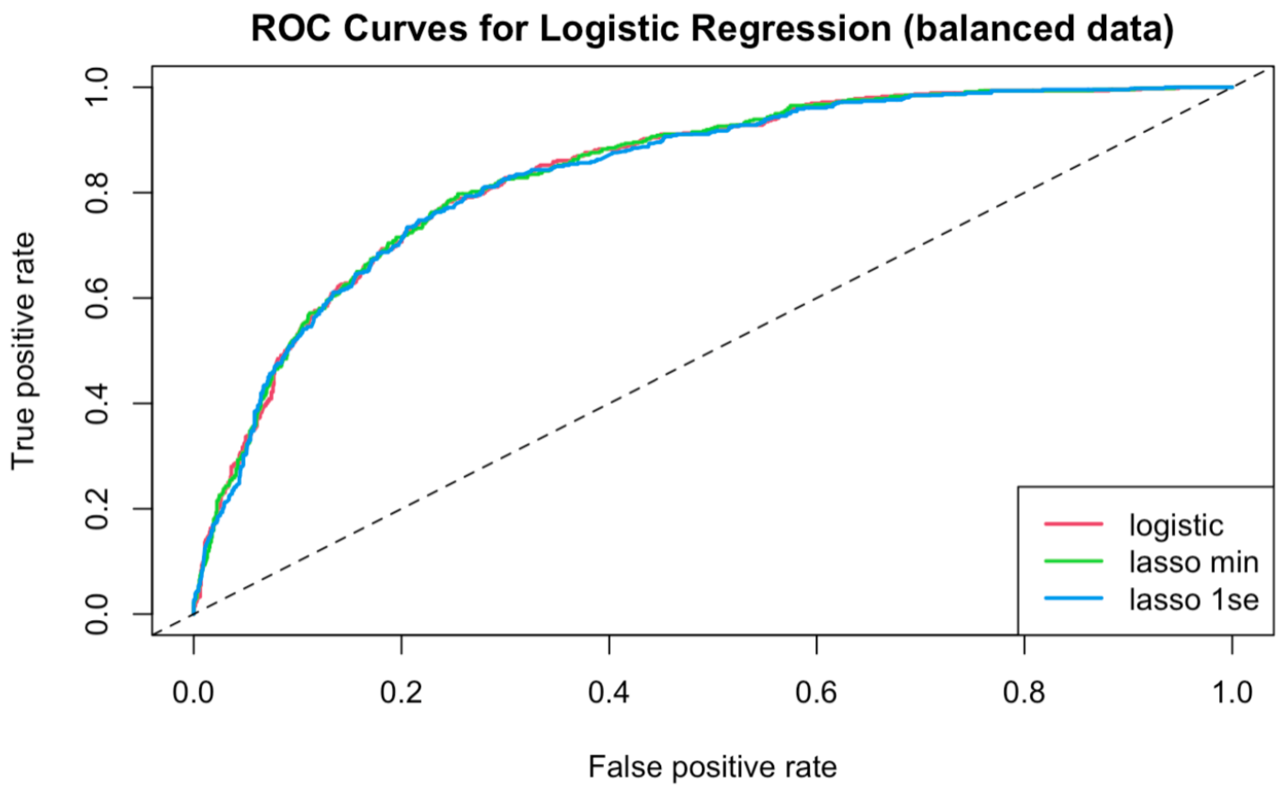
TELCO SYSTEMS

Logistic Regression – Balanced Data

- Regular logistic regression
- Lasso model with λ that minimizes cross-validated error
- Lasso model with λ in which the error is within 1 se of the minimum

➢ similar predictive powers in terms of deviance, accuracy, and AUC

➢ Compared to unbalanced data, the performance is worse off



CHICAGO BOOTH

TELCO SYSTEMS

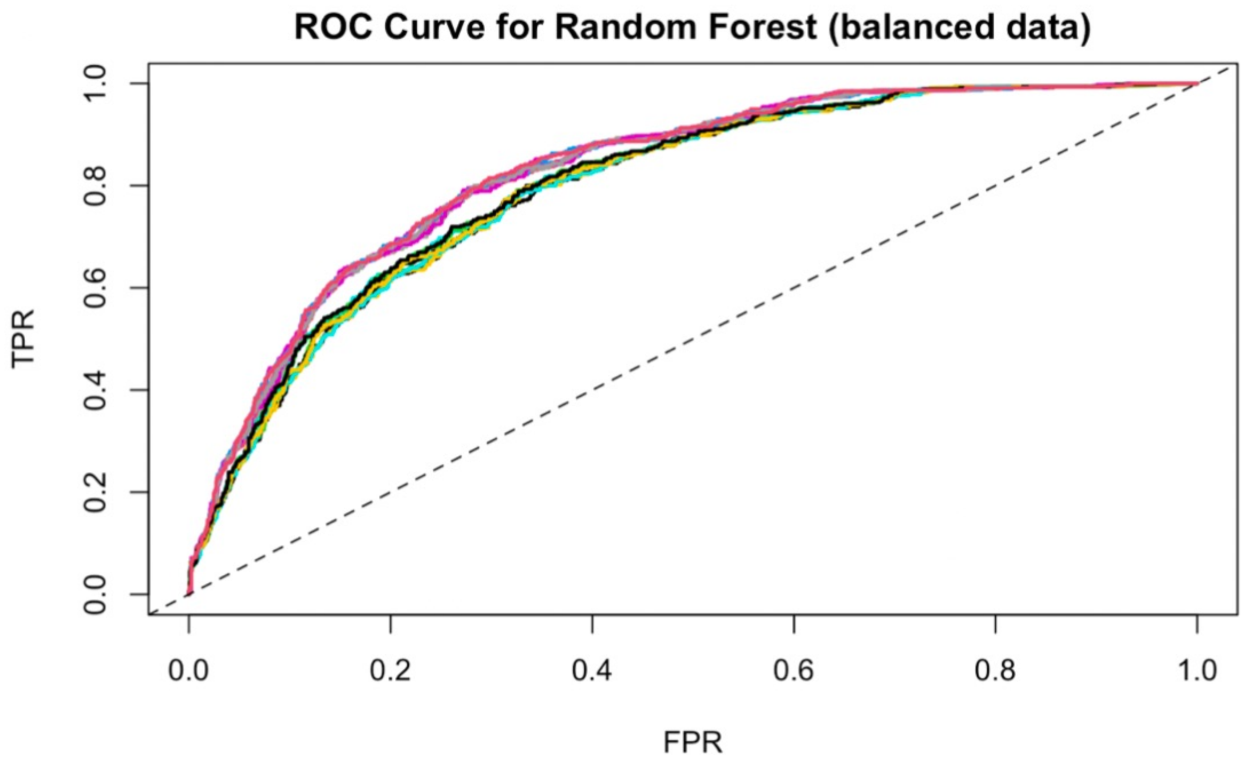
Random Forest – Balanced Data

Parameter search:

- number of trees
- number of variables to possibly split at in each node
- minimal node size to tune the parameters

Optimal parameters:

- ➢number of trees = 1000
- ➢number of variables to split = 5
- ➢ minimal node size = 20

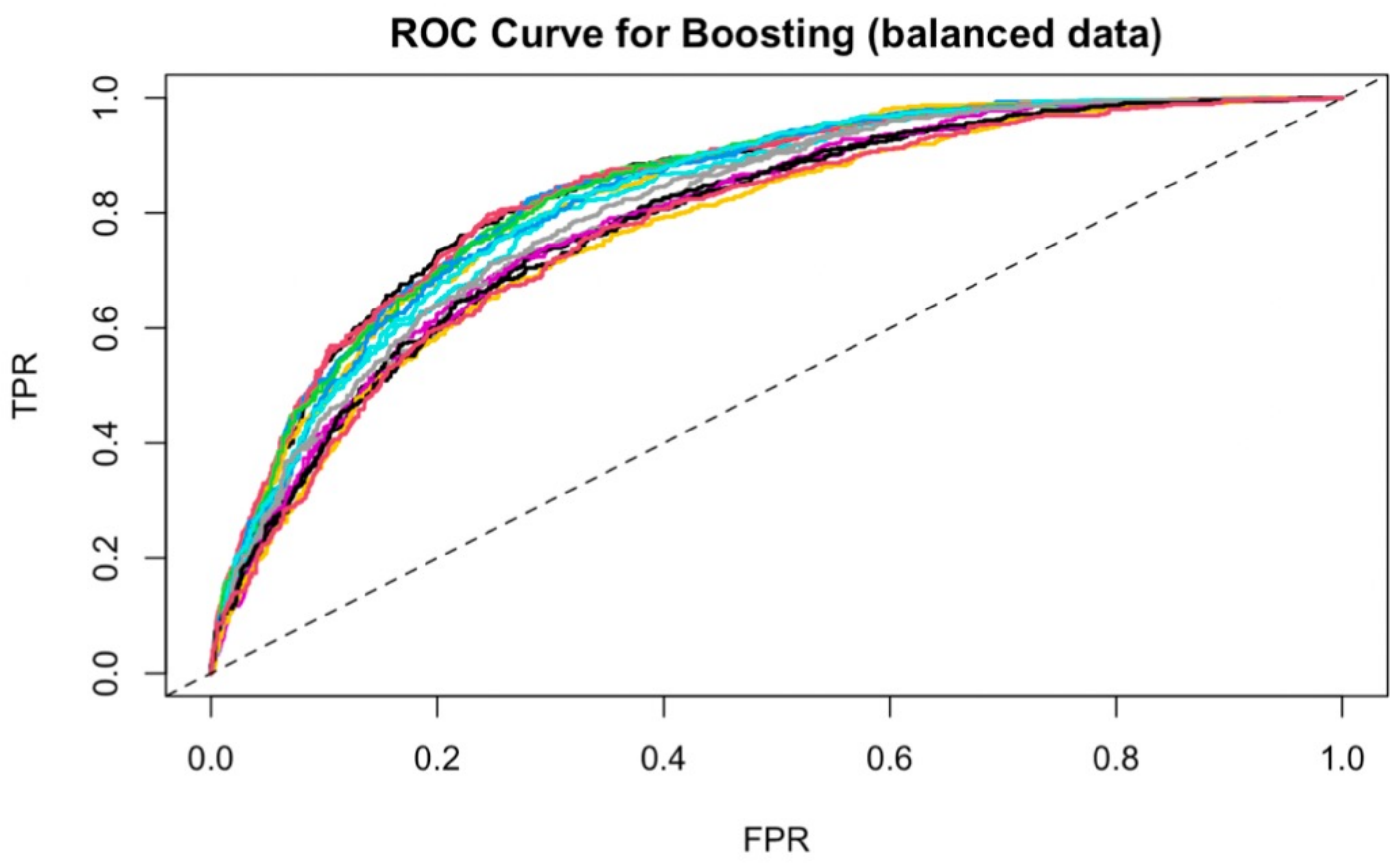


Boosting – Balanced Data

Parameter search:
- learning rate
- maximum depth of trees

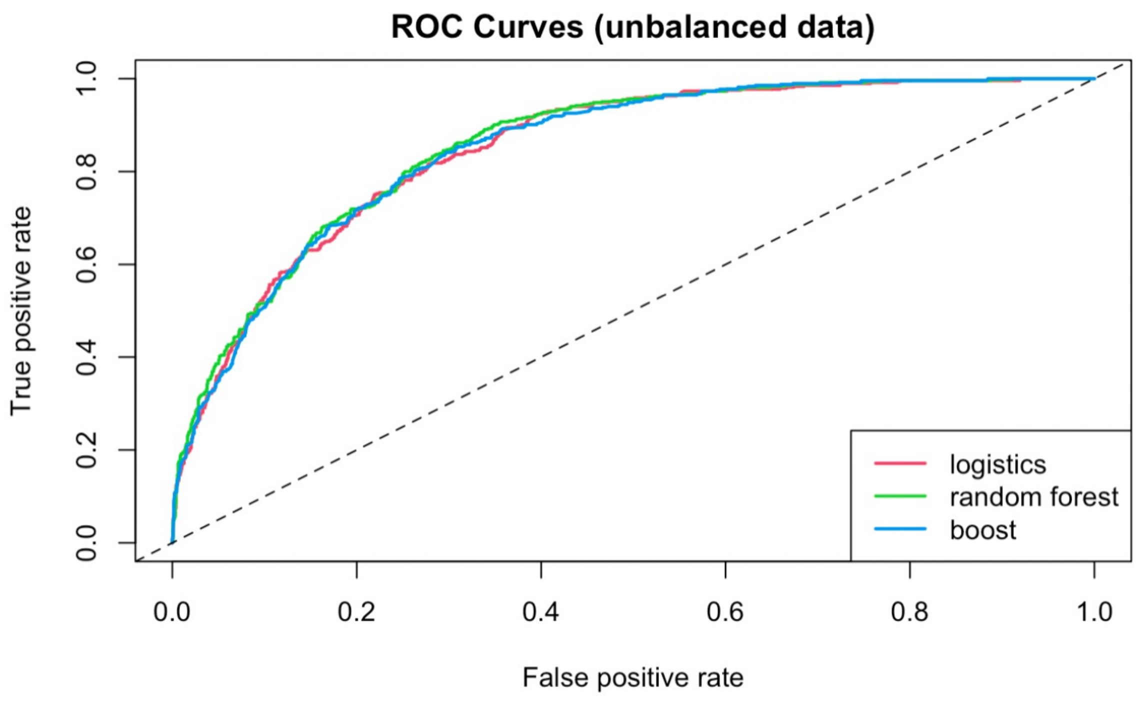
Optimal parameters:
- learning rate = 0.3
- maximum depth of trees = 5

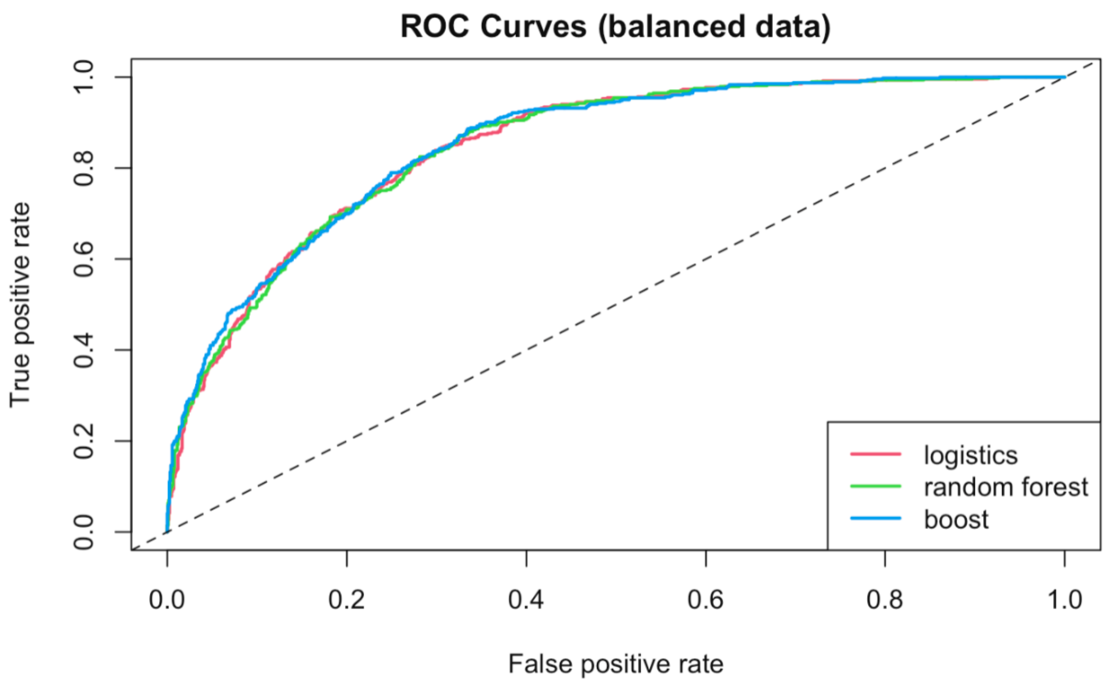
- Compared to unbalanced data, the performance of boosting trees improves substantially



Results – Unbalanced & Balanced Data

Random Forest is optimal for both





Results – Unbalanced & Balanced Data

Random Forest is optimal for both

A potential problem is, while the model is doing generally well in predicting users who stay, it doesn't predict well in users who churn

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1165  246
       Yes  108  239

               Accuracy : 0.7986
                 95% CI : (0.7791, 0.8172
    No Information Rate : 0.7241
    P-Value [Acc > NIR] : 3.318e-13
```

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1087  183
       Yes  186  302

               Accuracy : 0.7901
                 95% CI : (0.7703, 0.8089)
    No Information Rate : 0.7241
    P-Value [Acc > NIR] : 1.198e-10
```

CHICAGO BOOTH

TELCO SYSTEMS

Conclusions and Improvements

- Best-performing model: random forest
- 80% accuracy overall

- Issue faced: unbalanced dataset
- Using up-sampling but not successfully solving the issue
- Future improvement: focusing on better dealing with class imbalance

CHICAGO BOOTH

TELCO SYSTEMS