STOP THE CHURN!

Aaron McMoran

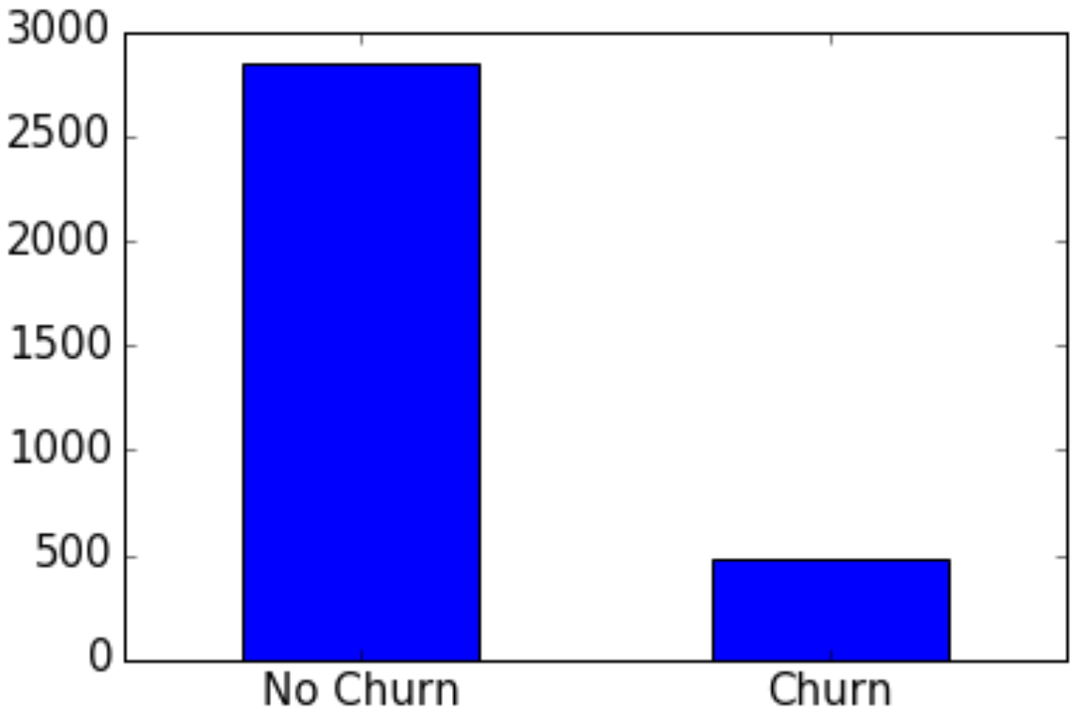
July 14th, 2016

What is churn and why do we care?

"The churn rate is the percentage of subscribers to a service who discontinue their subscriptions to that service within a given time period. " - Investopedia

Source

The Data



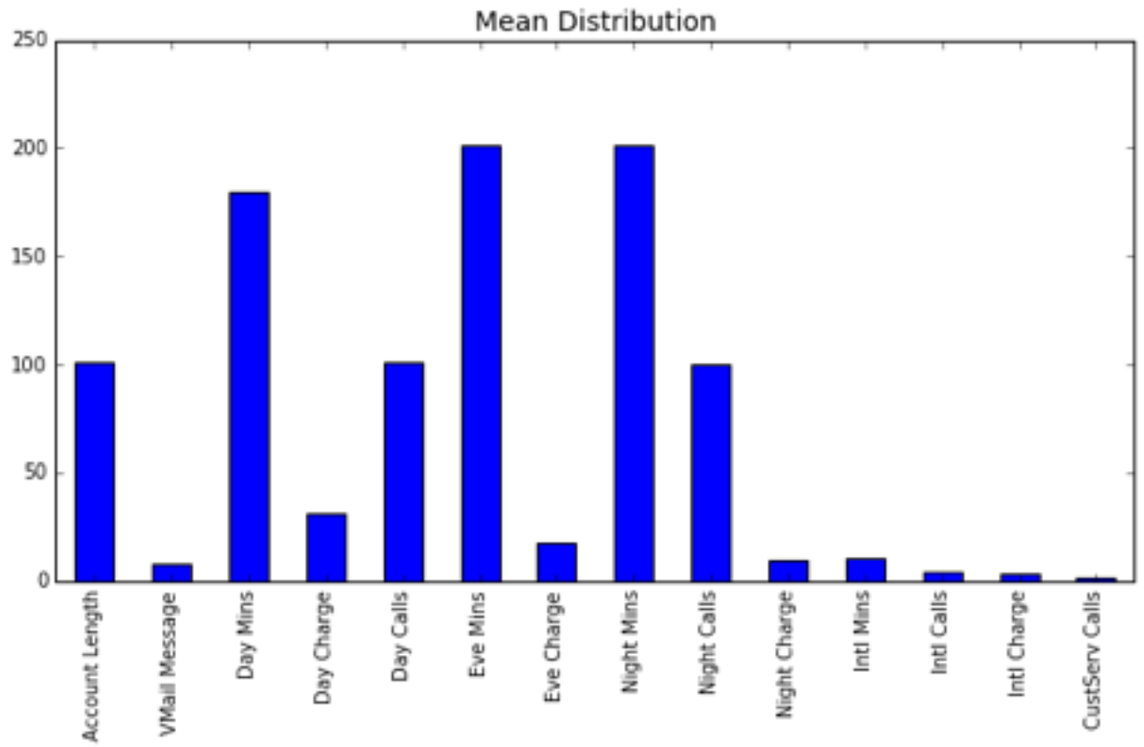
Imbalanced data!

```
State : object
Account Length : int64
Area Code : int64
Phone : object
Int'l Plan : object
VMail Plan : object
VMail Message : int64
Day Mins : float64
Day Calls : int64
Day Charge : float64
Eve Mins : float64
Eve Calls : int64
Eve Charge : float64
Night Mins : float64
Night Calls : int64
Night Charge : float64
Intl Mins : float64
Intl Calls : int64
Intl Charge : float64
CustServ Calls : int64
Churn? : object
```

4 Discrete features

16 Continuous features

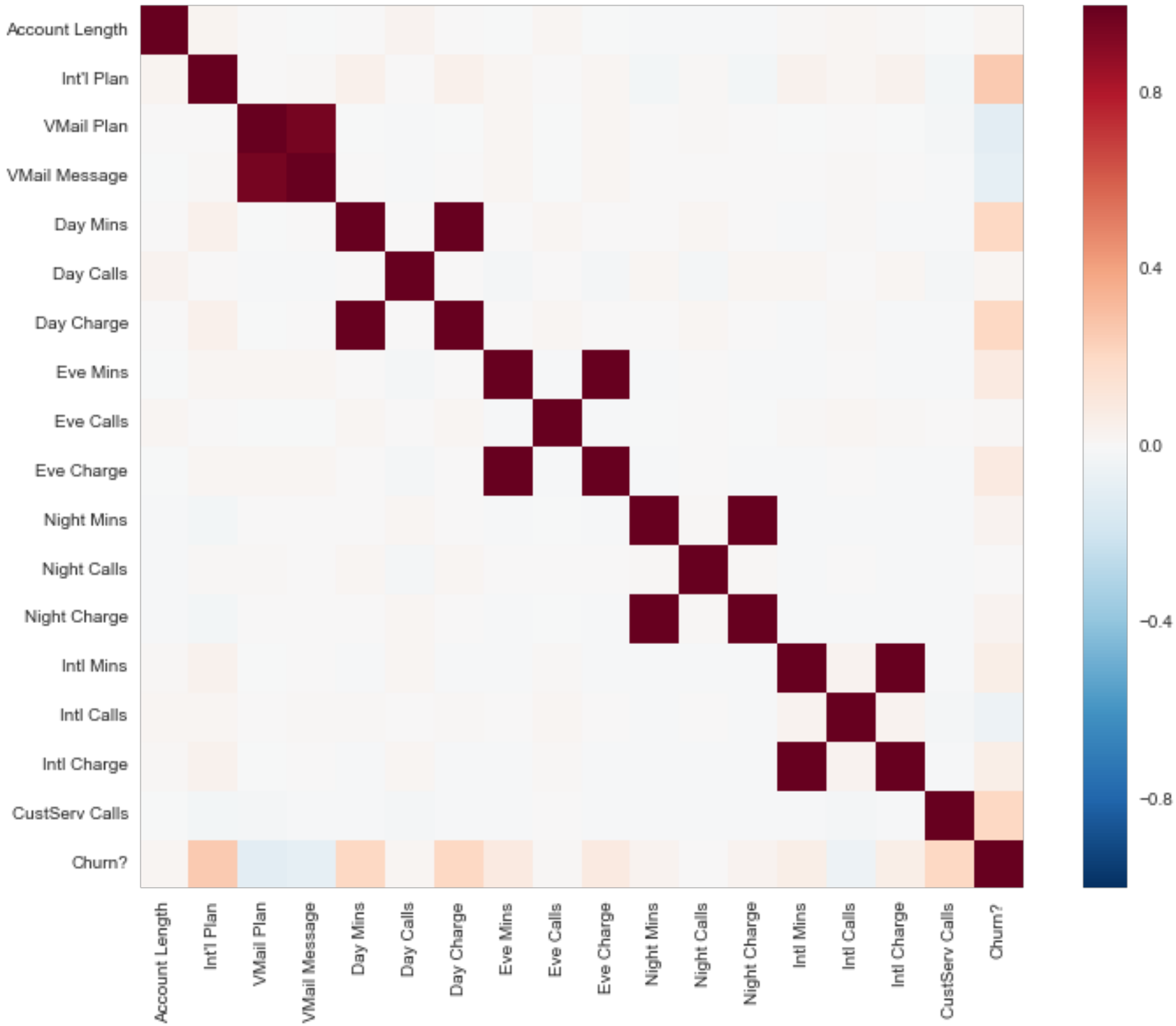
Data Cleaning



- Means of different magnitude indicate standardizing the data
- Drop unrelated features and turn categorical data to binary

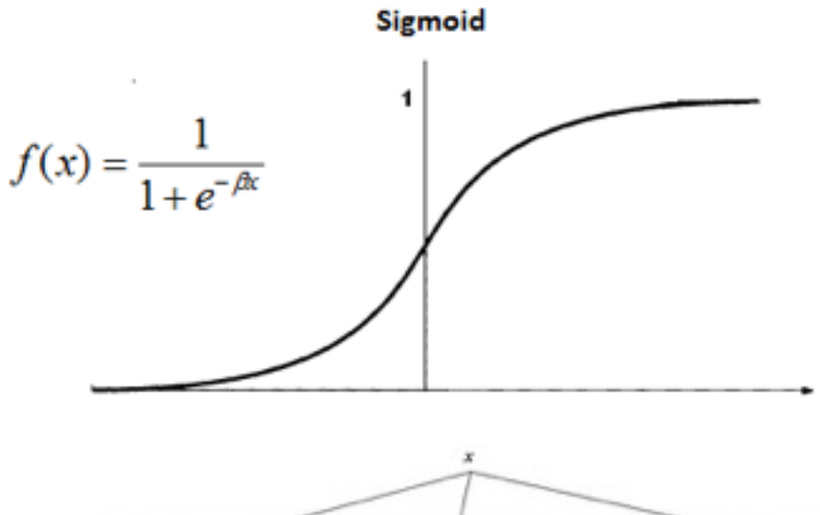
```
#Converting binary data
clean_churn["Int'l Plan"] = np.where(clean_churn["Int'l Plan"]=="yes", 1.0, 0.0)
clean_churn["VMail Plan"] = np.where(clean_churn["VMail Plan"]=="yes", 1.0, 0.0)
clean_churn["Churn?"] = np.where(clean_churn["Churn?"]=="True.", 1.0, 0.0)
#Dropping noisy data
clean_churn = churn_data.drop(['Area Code','State', 'Phone'], axis = 1)
```

Correlation Matrix & Collinear Features

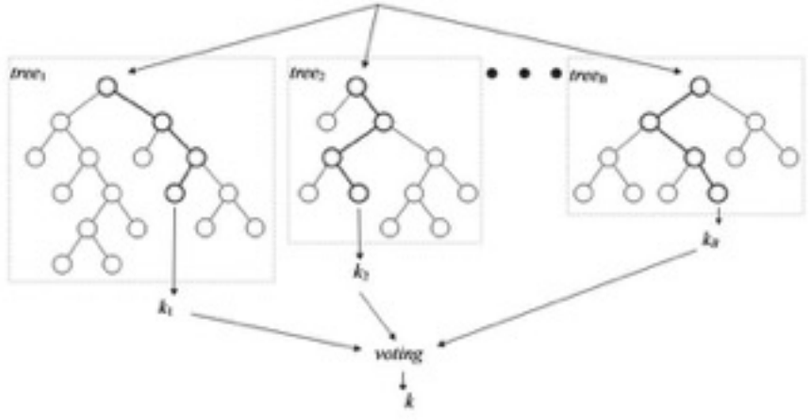


Models

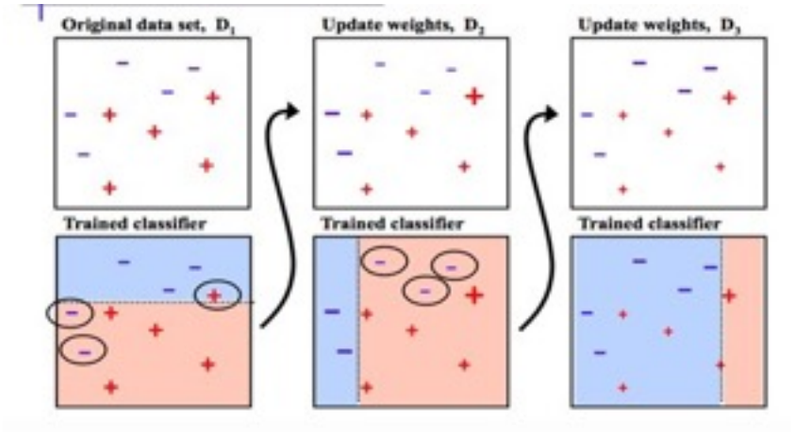
Logistic Regression



Random Forest



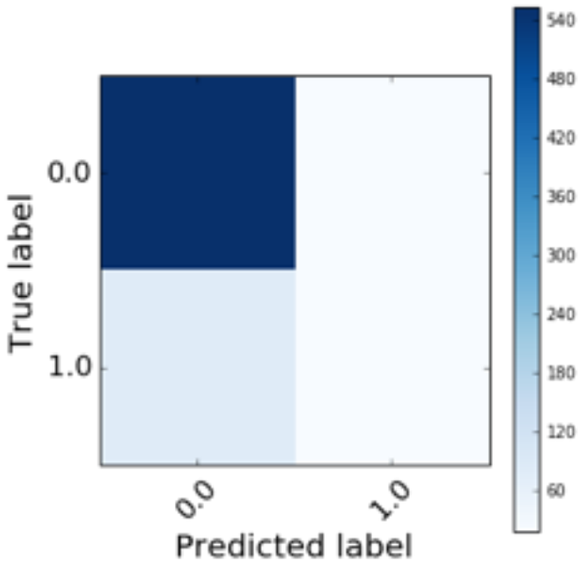
AdaBoost



Results - Confusion Matrix

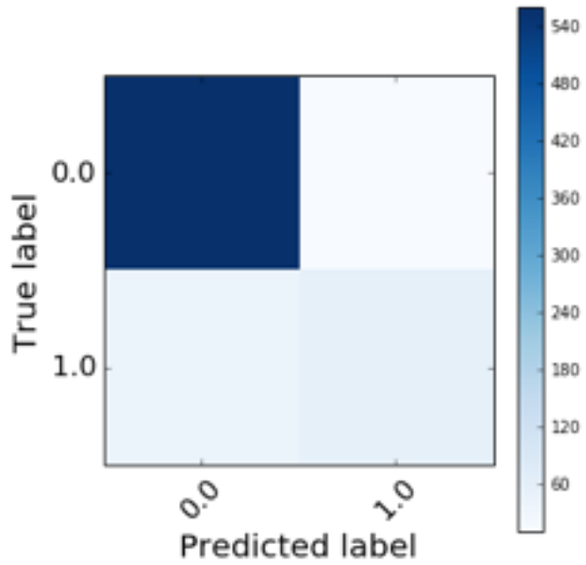
Winner!

Logistic regression



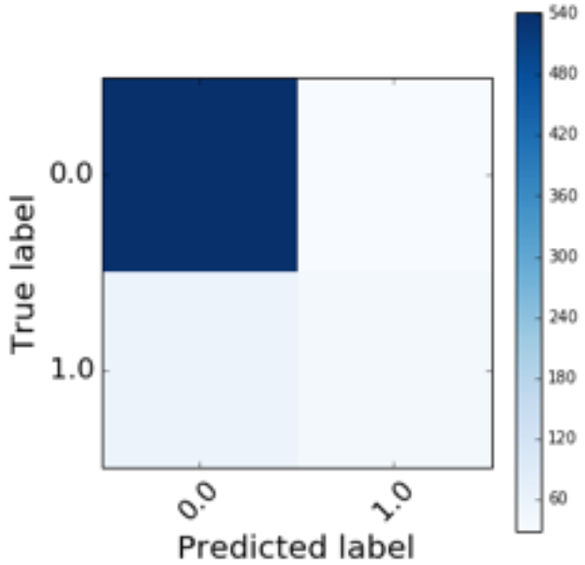
	Predicted Label	
True Label	552	17
	81	17

Random Forest



	Predicted Label	
True Label	560	9
	42	56

AdaBoost

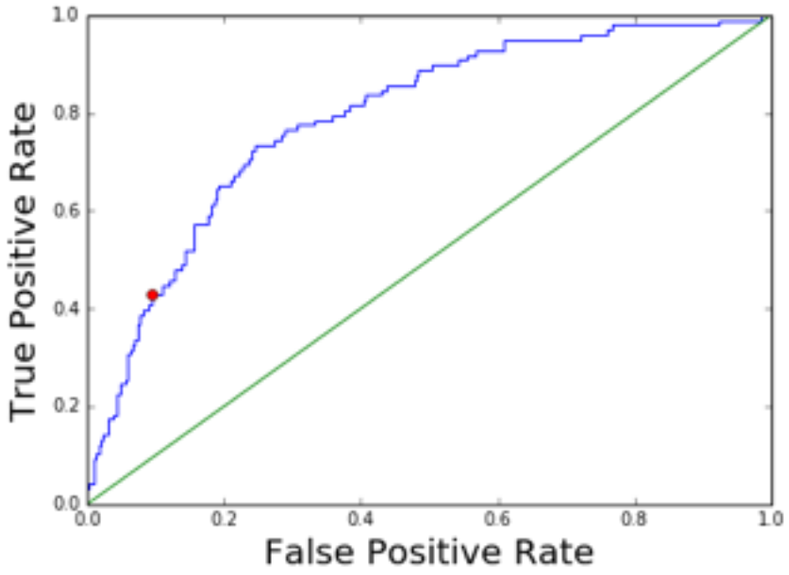


	Predicted Label	
True Label	541	28
	58	40

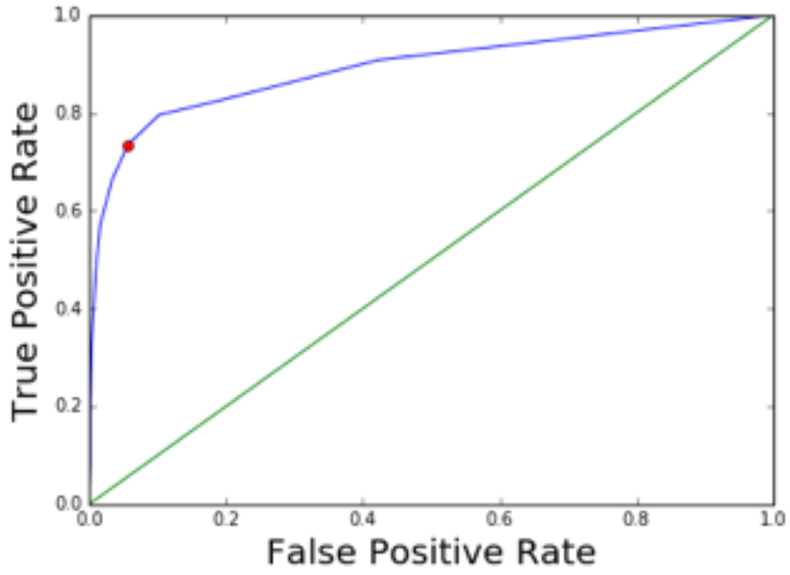
Results - Roc AUC Curve

Winner!

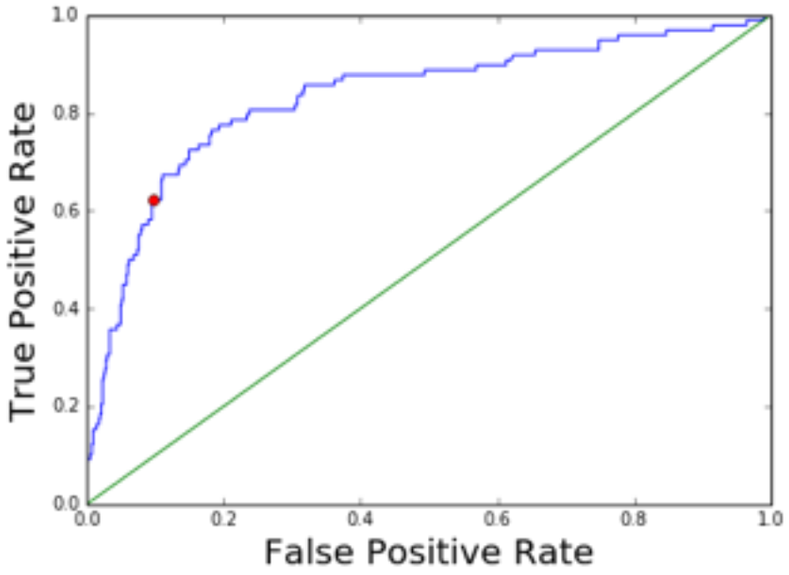
Logistic Regression



Random Forest



AdaBoost

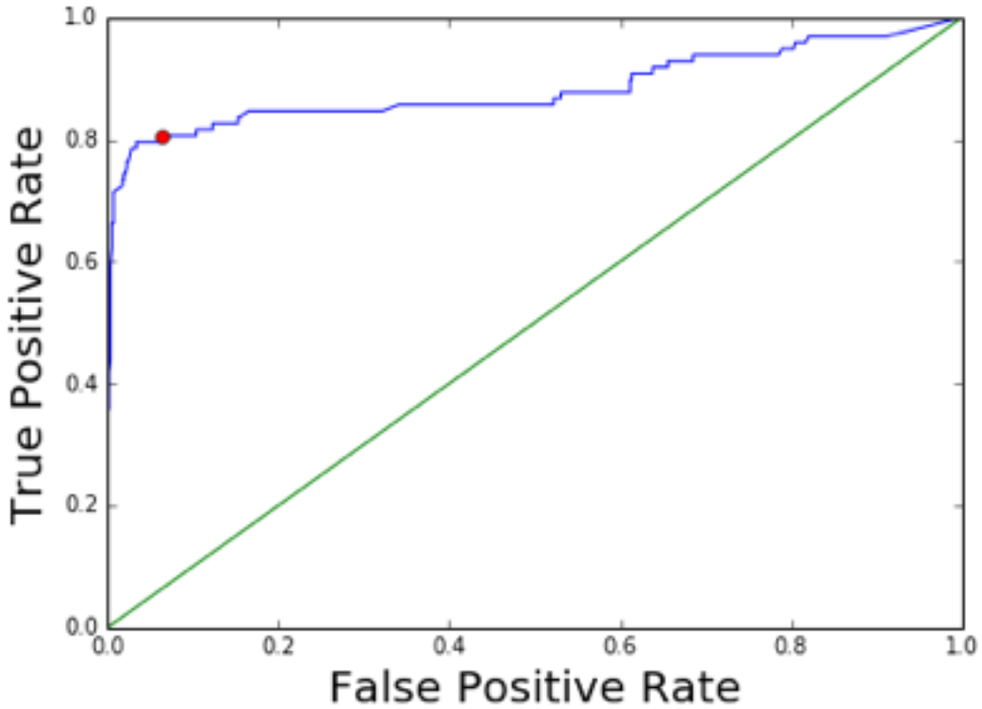


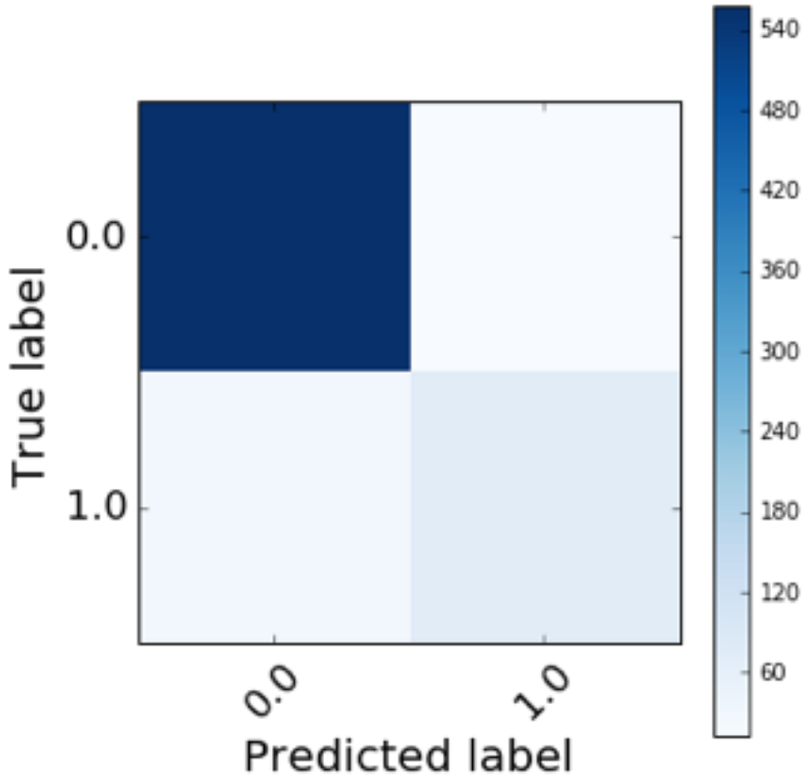
	FPR @ max 10% FPR	TPR @ max 10% FPR	Mean Roc AUC	5- Kfold Variance
Logistic Regression	9.4%	42.9%	82%	2%%
Random Forrest	**5.6%**	**73.5%**	**91%**	**3%%**
AdaBoost	9.7%	62.2%	87%	2%%

Tuning Hyper-parameters for Random Forest

- Number of estimators
- Criterion
- Maximum Features
- Class weight
- Minimum Sample Split

Optimized Random Forest

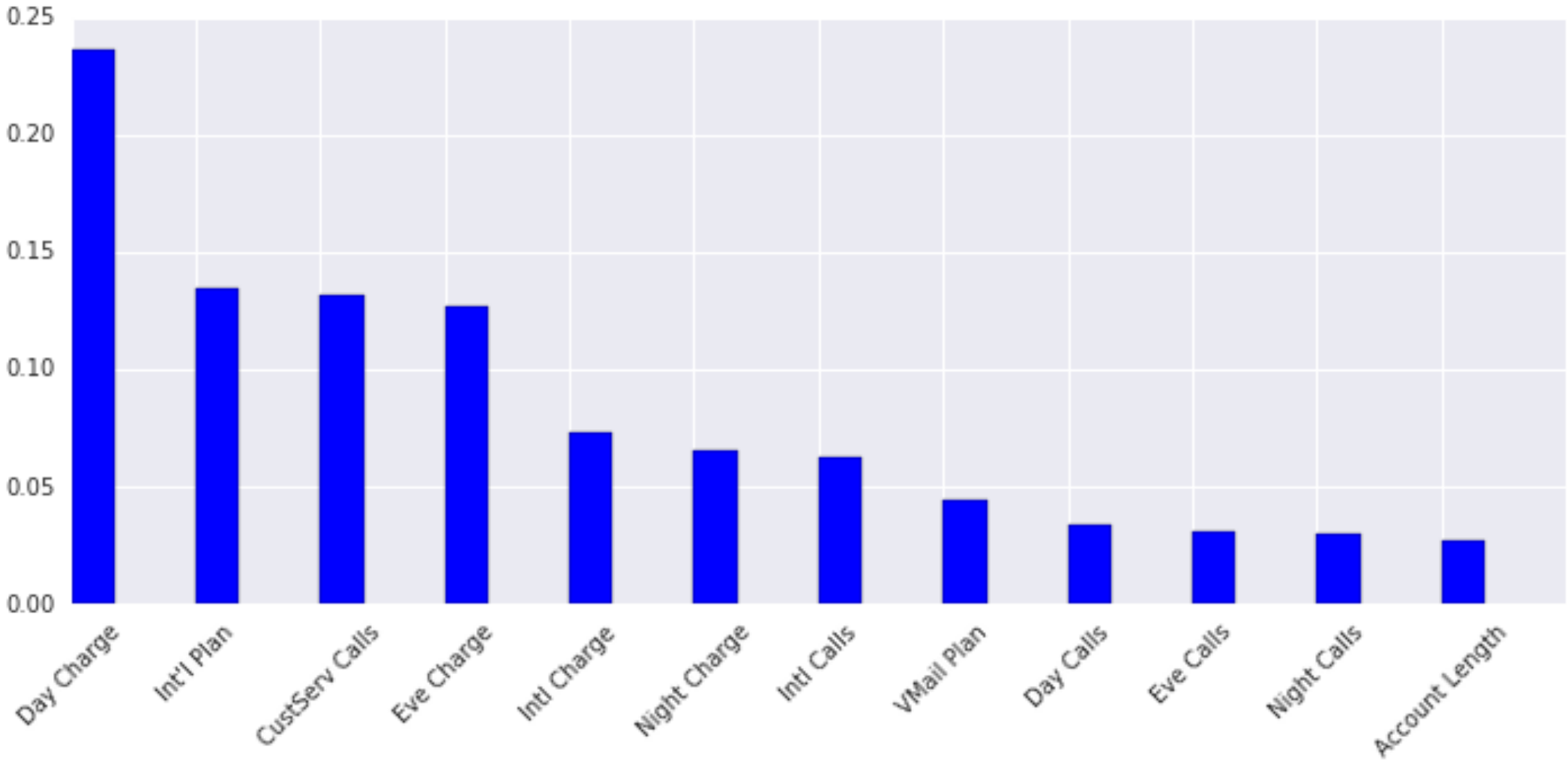




	FPR @ max 10% FPR	TPR @ max 10% FPR	Mean ROC AUC	5 Kfold Variance
Tuned Random Forest	**6.3%**	**80.6%**	**94%**	**3%**

	Predicted Label	
True Label	557	12
	25	73

Which features do we care about?



THANK YOU!