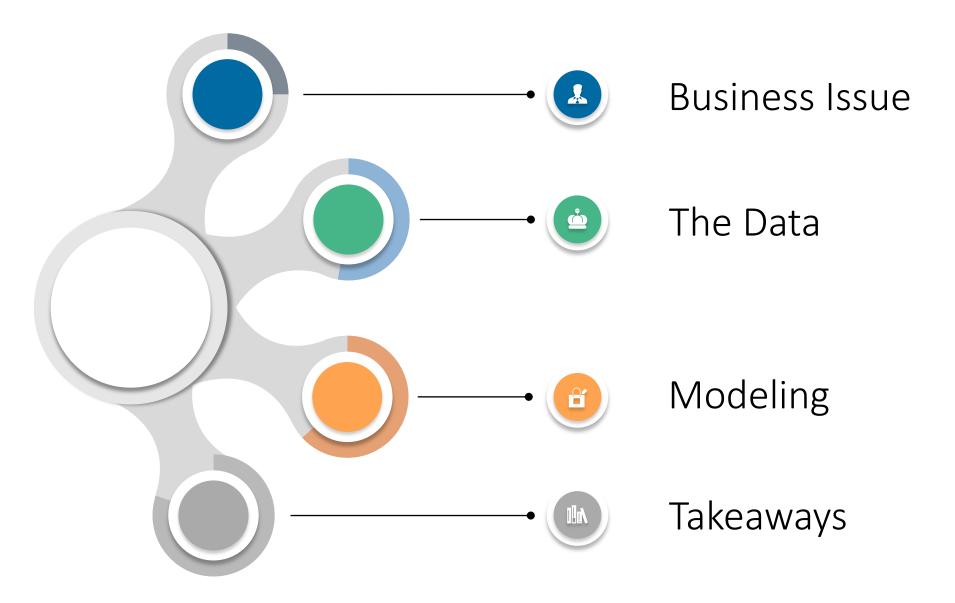


Supervised Learning Capstone

Will Becerra

# Agenda



### What is churn?



Churn: when a customer leaves the service of one company to become a customer of another company.

# Challenges in the Telco Industry

Market saturation

Product lifecycles shortening



Increased competition

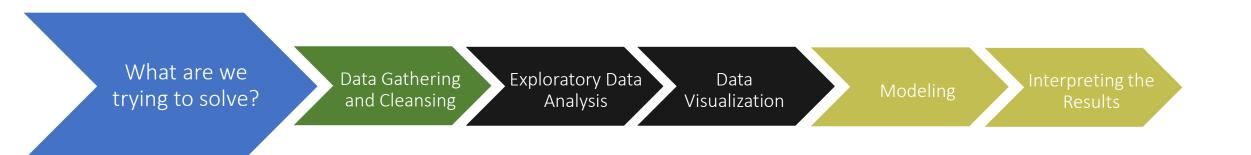
Low differentiation

Disruption from OTT players

# Telecom Industry Averages

- Churn rate is 1.9% (20.5% yearly) across top four carriers (AT&T, Verizon, T-Mobile, Sprint)
- Customer base is 100M across top four carriers
- Monthly loss from churn, per carrier is \$65M
- Acquisition cost for a new customer is \$315
- Lost revenue from churned customer is \$1,117
- Customer lifetime value us \$1,782

# Objective and Approach



Where can a company have the biggest impact in reducing churn?

# Summary of Findings

- Having a plan matters (International, Voice Mail).
- Usage charges and customer service calls impact churn.
- The accuracy level reached was 94.3%.
- Best models were Random Forest and Gradient Boosting.

#### Data Set & Features

- 3333 customers
- 20 predictors (features)

#### **Predict Binary Classification**

 Churn (target variable): whether the customer left the service: true/false

- State
- Area Code
- Phone
- Int'l Plan: yes/no
- VMail Plan: yes/no
- VMail Message: voice mails / per month
- Day Mins, Day Calls, Day Charge
- Eve Mins, Eve Calls, Eve Charge
- Night Mins, Night Calls, Night Charge
- Intl Mins, Intl Calls, Intl Charge
- CustServ Calls
- Account Length: tenure with the company

# Features at a glance

- The International Plan and Voice Mail Plans are categorical, either yes or no.
- The phone number is a unique customer identifier.
- The remaining columns are either integers or continuous values.
- The target is our churn variable, either true or false.
- There are no missing values.

# **Exploratory Data Analysis**



account_length	1	0.025	0.0029	0.0046	0.0062	0,038	0.0062	-0.0068	0.019	-0.0067	-0.009	-0.013	-0.009	0.0095	0.021	0.0095	-0.0038	0.017
intl_plan	0.025	1	0.006	0.0087	0.049	0.0038	0.049	0.019	0.0061	0.019	-0.029	0.012	-0.029	0.046	0.017	0.046	-0.025	0.26
vmail_plan	0.0029	0.006	1	0.96	-0.0017	-0.011	-0.0017	0.022	-0.0064	0.022	0.0061	0.016	0.0061	-0.0013	0.0076	-0.0013	-0.018	-0.1
vmail_message	-0.0046	0.0087	0.96	1	0.00078	-0.0095	0.00078	0.018	-0.0059	0.018	0.0077	0.0071	0.0077	0.0029	0.014	0.0029	-0.013	-0.09
day_mins	0.0062	0.049	-0.0017	0.00078	1	0.0068	10	0.007	0.016	0.007	0.0043	0.023	0.0043	-0.01	0.008	-0.01	-0.013	0.21
day_calls	0.038	0.0038	-0.011	-0.0095	0.0068	1	0.0068	-0.021	0.0065	-0.021	0.023	-0.02	0.023	0.022	0.0046	0.022	-0.019	0.018
day_charge	0.0062	0.049	-0.0017	0.00078	1	0.0068	1.	0.007	0.016	0.007	0.0043	0.023	0.0043	-0.01	0.008	-0.01	-0.013	0.21
eve_mins	-0.0068	0.019	0.022	0.018	0.007	-0.021	0.007	1	-0.011	1	-0.013	0.0076	-0.013	-0.011	0.0025	-0.011	-0.013	0.093
eve_calls	0.019	0.0061	-0.0064	-0.0059	0.016	0.0065	0.016	-0.011	1	-0.011	-0.0021	0.0077	-0.0021	0.0087	0.017	0.0087	0.0024	0.0092
eve_charge	-0.0067	0.019	0.022	0.018	0.007	-0.021	0.007	1	-0.011	1	-0.013	0.0076	-0.013	-0.011	0.0025	-0.011	-0.013	0.093
night_mins	-0.009	-0.029	0.0061	0.0077	0.0043	0.023	0.0043	-0.013	-0.0021	-0.013	1	0.011	1	-0.015	-0.012	-0.015	-0.0093	0.035
night_calls	-0.013	0.012	0.016	0.0071	0.023	-0.02	0.023	0.0076	0.0077	0.0076	0.011	1	0.011	-0.014	0.0003	-0.014	-0.013	0.0061
night_charge	-0.009	-0.029	0.0061	0.0077	0.0043	0.023	0.0043	-0.013	-0.0021	-0.013	1	0.011	1	-0.015	-0.012	-0.015	-0.0093	0.035
intl_mins	0.0095	0.046	0.0013	0.0029	-0.01	0.022	-0.01	-0.011	0.0087	-0.011	-0.015	-0.014	-0.015	1	0.032	1	-0.0096	0.068
intl_calls	0.021	0.017	0.0076	0.014	0.008	0.0046	0.008	0.0025	0.017	0.0025	-0.012	0.0003	-0.012	0.032	1	0.032	-0.018	-0.053
intl_charge	0.0095	0.046	-0.0013	0.0029	-0.01	0.022	-0.01	-0.011	0.0087	-0.011	-0.015	-0.014	-0.015	1	0.032	1	-0.0097	0.068
custserv_calls	0.0038	-0.025	-0.018	-0.013	-0.013	-0.019	-0.013	-0.013	0.0024	-0.013	-0.0093	-0.013	-0.0093	-0.0096	-0.018	-0.0097	12	0.21
churn	0.017	0.26	-0,1	-0.09	0.21	0.018	0.21	0.093	0.0092	0.093	0.035	0.0061	0.035	0.068	-0.053	0.068	0.21	18
	account_length	intl_plan	wmail_plan	vmail_message	day_mins	day_calls	day_charge	eve_mins	eve_calls	eve_charge	night_mins	night_calls	night_charge	intl_mins	intl_calls	intl_charge	custserv_calls	churn

Pearson Correlation

-0.2

# Dimensionality Reduction

- Removed highly correlated features to reduce the dimensions (no sense making the model do more work than necessary)
  - Day mins / day charges
  - Eve mins / eve charges
  - Night mins / night charges
  - International mins / International charges
- Removed area code
- Removed phone numbers
- Turned the state variable into 51 indicator variables (value 0 or 1)

account_length ·	1	0.025	0.0029	-0.0046	0.038	0.0062	0.019	-0.0067	-0.013	-0.009	0.021	0.0095	-0.0038	0.017
intl_plan -	0.025	i	0.006	0.0087	0.0038	0.049	0.0061	0.019	0.012	-0.029	0.017	0.046	-0.025	0.26
vmail_plan -	0.0029	0.006	1	0.96	-0.011	0.0017	-0.0064	0.022	0.016	0.0061	0.0076	-0.0013	-0.018	-0.1
mail_message ·	-0.0046	0.0087	0.96	1	-0.0095	0.00078	-0.0059	0.018	0.0071	0.0077	0.014	0.0029	-0.013	-0.09
day_calls -	0.038	0.0038	-0.011	-0.0095	1	0.0068	0.0065	-0.021	-0.02	0.023	0.0046	0.022	-0:019	0.018
day_charge ·	0.0062	0.049	-0.0017	0.00078	0.0068	1	0.016	0.007	0.023	0.0043	0.008	-0.01	-0.013	0.21
eve_calls ·	0.019	0.0061	-0.0064	-0.0059	0.0065	0.016	1	-0.011	0.0077	-0.0021	0.017	0.0087	0.0024	0.0092
eve_charge -	-0.0067	0.019	0.022	0.018	-0.021	0.007	-0.011	1	0.0076	-0.013	0.0025	-0.011	-0.013	0.093
night_calls -	-0.013	0.012	0.016	0.0071	-0.02	0.023	0.0077	0.0076	1	0.011	0.0003	-0.014	-0.013	0.0063
night_charge	-0.009	-0.029	0.0061	0.0077	0.023	0.0043	-0.0021	-0.013	0.011	1	-0.012	-0.015	-0.0093	0.035
intl_calls -	0.021	0.017	0.0076	0.014	0.0046	0.008	0.017	0.0025	0.0003	-0.012	1	0.032	-0.018	-0.053
intl_charge -	0.0095	0.046	-0.0013	0.0029	0.022	-0.01	0.0087	-0.011	-0.014	-0.015	0.032	1	-0.0097	0.068
oustserv_calls ·	-0.0038	-0.025	-0.018	-0.013	-0.019	-0.013	0.0024	-0.013	-0.013	-0.0093	-0.018	-0.0097	1	0.21
churn -	0.017	0.26	-0.1	-0.09	0.018	0.21	0.0092	0.093	0.0061	0.035	-0.053	0.068	0.21	1
	- utgnai	tl plan -	il plan -	- ačessa	y_calls -	charge -	e_calls -	charge -	nt_calls -	charge -	ti calls -	charge -	v calls -	dum -

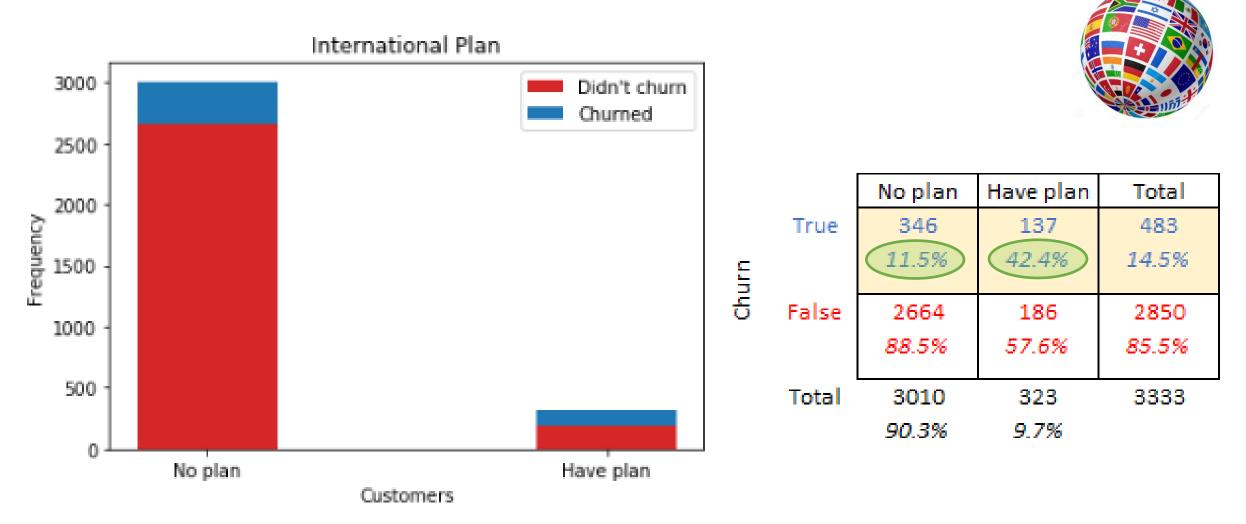
# Correlation with churn

Intl Plan: 0.26

Day charge: 0.21

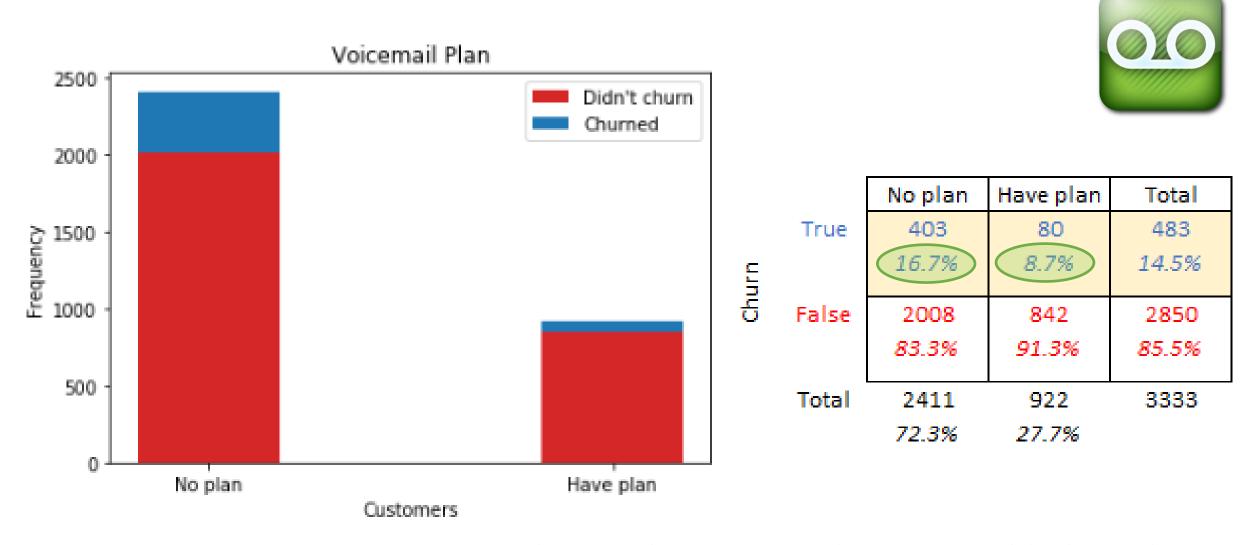
Customer service calls: 0.21

### Customers with International Plan



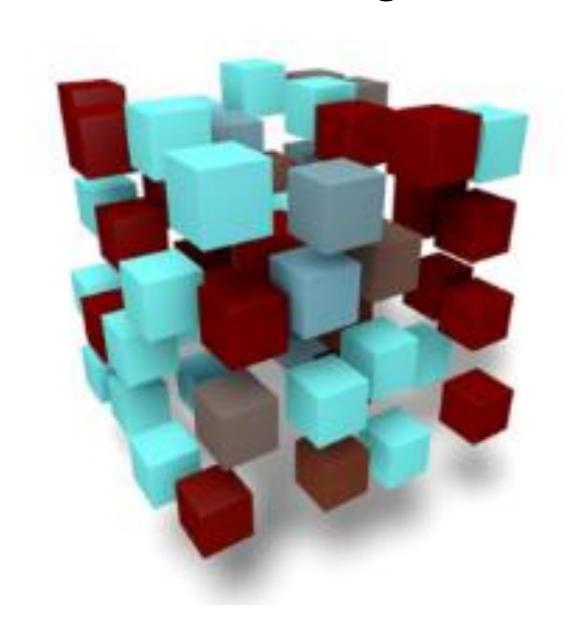
With plan, more than 3X as likely to leave

### Customers with Voicemail Plan

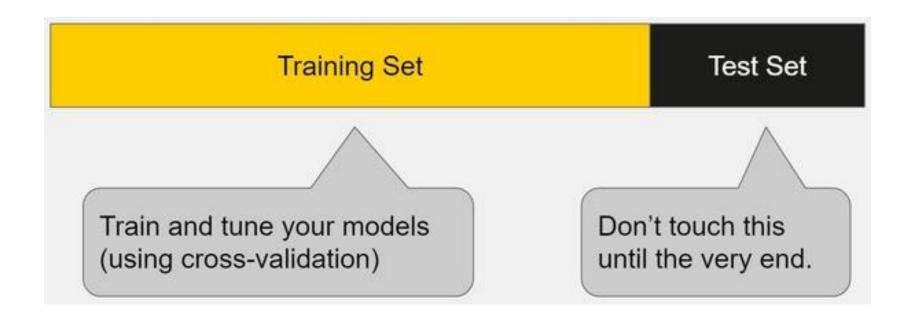


Without plan, more than 2X as likely to leave

# Modeling



### Divided data set train/test - 20%



### Dealing with Imbalance of Classes

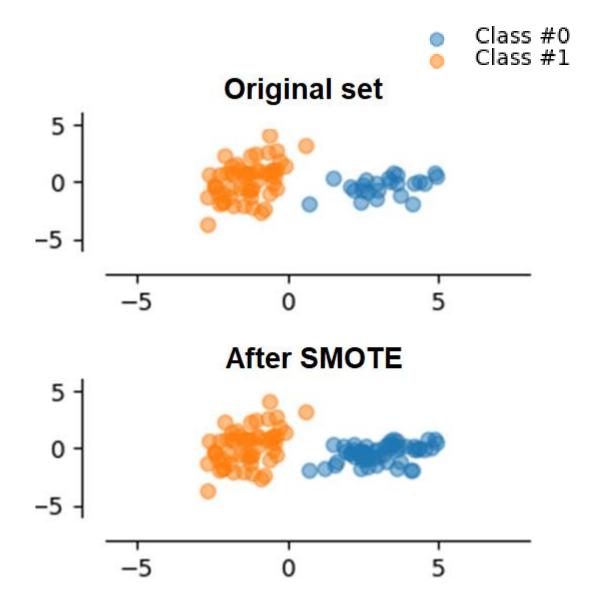
Only 14.9 % churned → class imbalance

Synthetic Minority Oversampling Technique (SMOTE)

```
sm = SMOTE(random_state=42)
X_train, y_train = sm.fit_sample(X_train, y_train)
print('Original dataset shape: {}'.format(Counter(Y)))
print()
print('Resampled dataset shape: {}'.format(Counter(y_train)))

Original dataset shape: Counter({0: 2850, 1: 483})

Resampled dataset shape: Counter({0: 2271, 1: 2271})
```



### Models

- Logistic Regression
- Ridge Regression
- Naïve Bayes
- KNN
- Decision Tree
- Random Forest
- Support Vector Machine
- Gradient Boosting



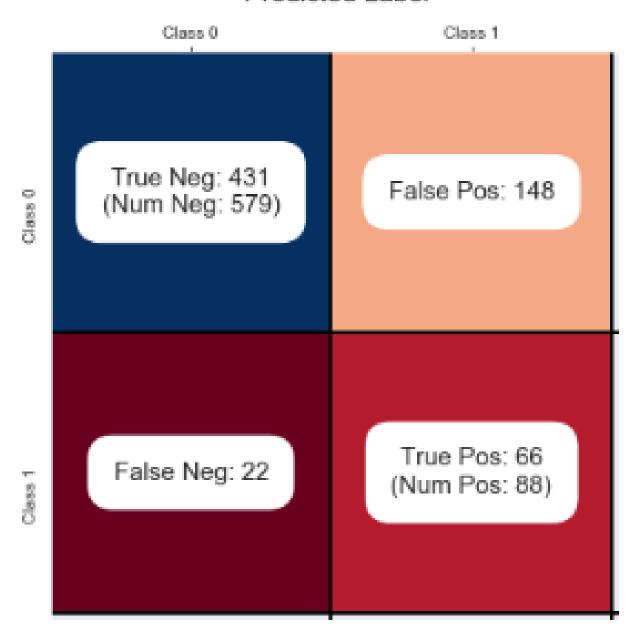
### What if we guessed that no one churned?

### Null Accuracy

```
# Null accuracy
max(y_test.mean(), 1- y_test.mean())
0.8680659670164917
```

Guess right 86.8% of the time

#### Predicted Label



When dealing with churn...

Goal: Reduce false negatives as much as possible

We don't want 'misses'



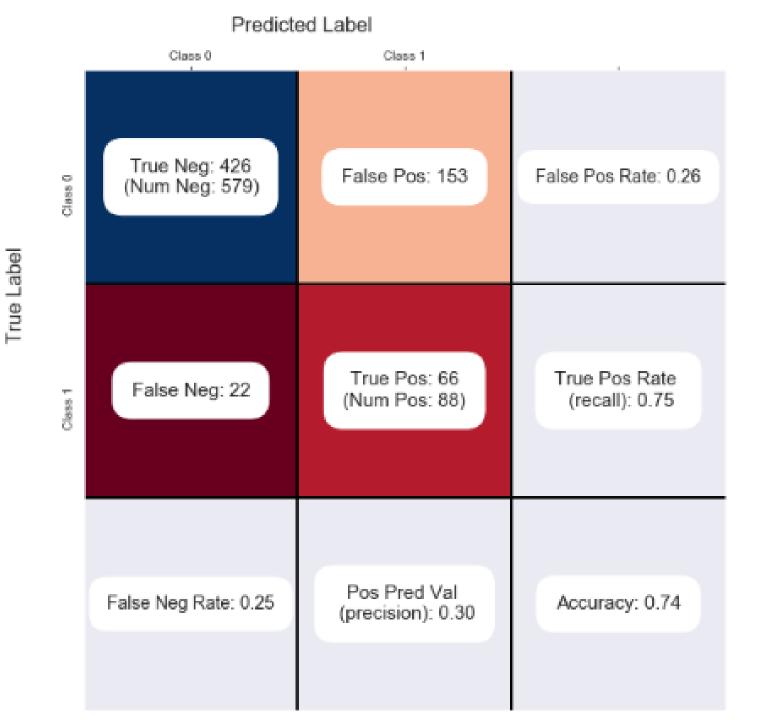
#### **Logistic Regression**

Accuracy: 0.745

Cross-validation: 0.793 +/- 0.011

Sensitivity 0.750 Ability of a model to find all the relevant cases within a dataset.

False negative rate: 0.25



#### **Ridge Classifier**

Accuracy: 0.738

Cross-validation: 0.792 +/- 0.011

False negative rate: 0.25

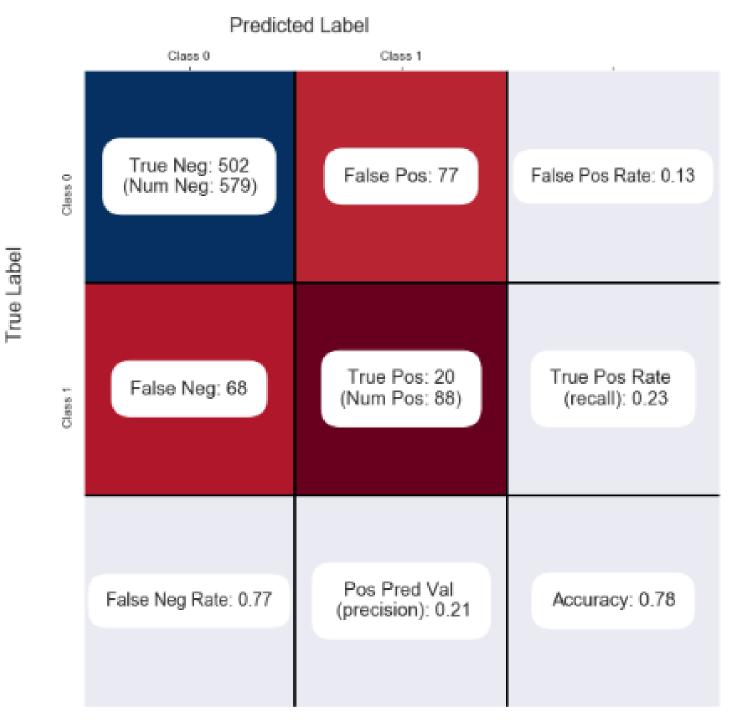
#### Predicted Label Class 1 Class 0 True Neg: 454 False Pos: 125 False Pos Rate: 0.22 (Num Neg: 579) True Pos: 38 True Pos Rate False Neg: 50 (Num Pos: 88) (recall): 0.43 Pos Pred Val False Neg Rate: 0.57 Accuracy: 0.74 (precision): 0.23

#### **Naïve Bayes**

Accuracy: 0.738

Cross-validation: 0.787 +/- 0.057

False negative rate: 0.57



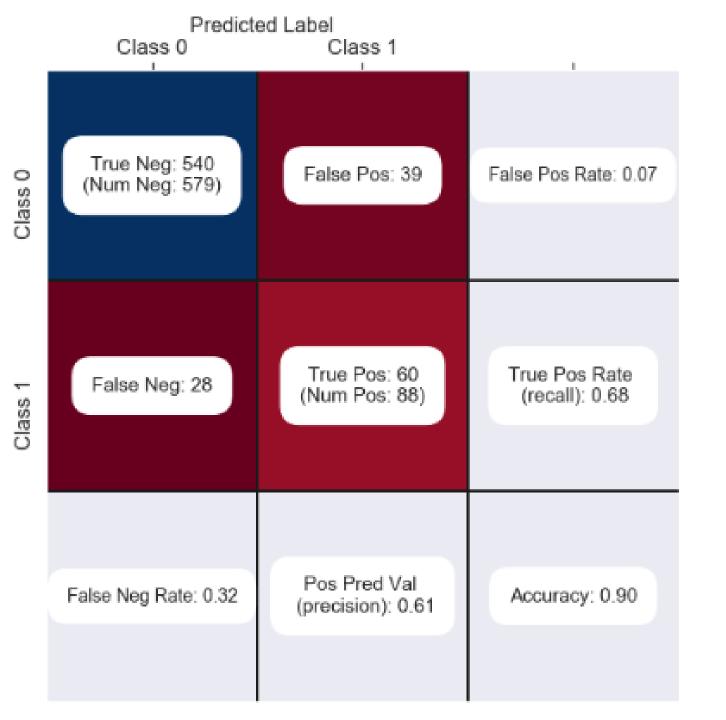
#### **K Nearest Neighbors**

Training accuracy: 0.994

Testing accuracy: 0.783

Difference = 0.211
Seriously overfitting

False negative rate: 0.77

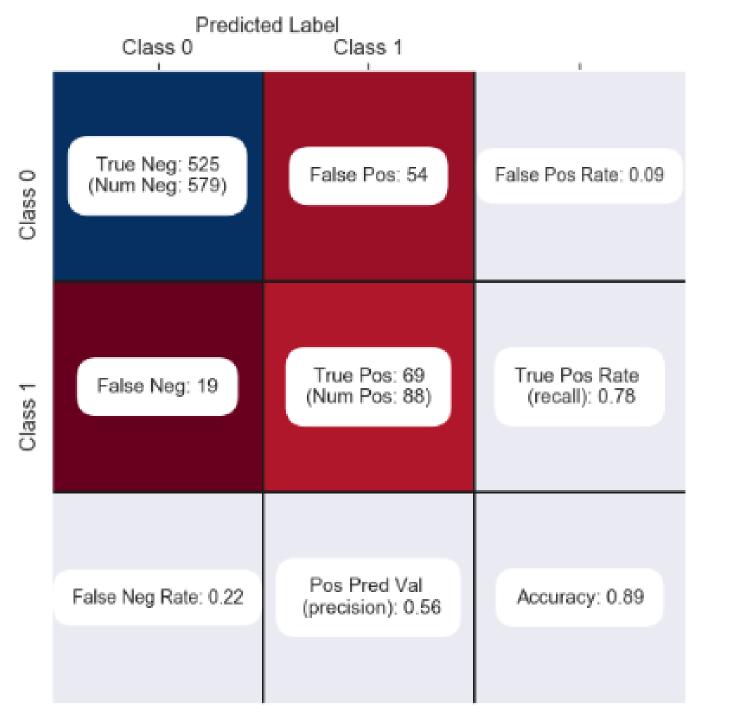


#### **Decision Tree**

Accuracy: 0.900

Cross-validation: 0.841 +/- 0.031

False negative rate: 0.32

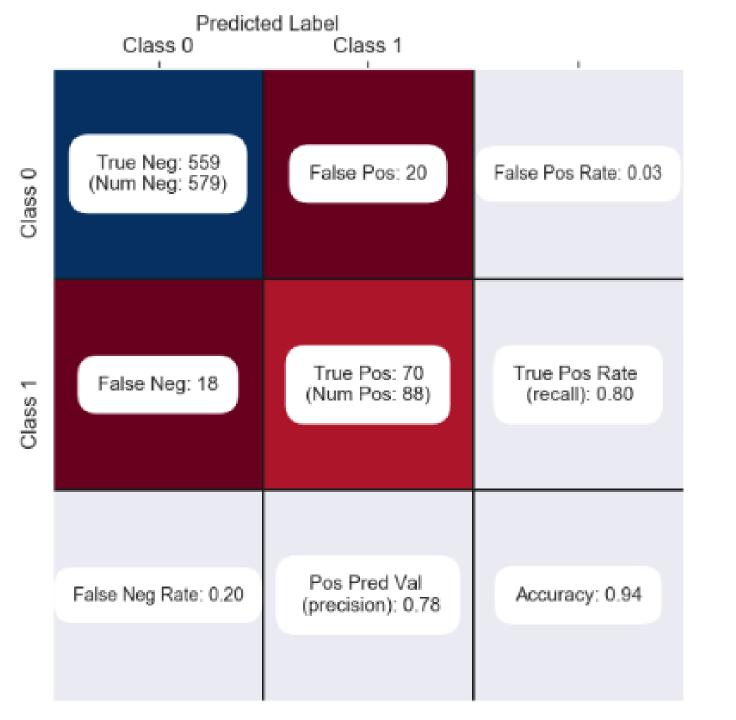


#### **Random Forest**

Accuracy: 0.891

Cross-validation: 0.918 +/- 0.026

False negative rate: 0.22



#### **Gradient Boosting**

Accuracy: 0.943

Cross-validation: 0.937 +/- 0.077

False negative rate: 0.20

(Best Scores)

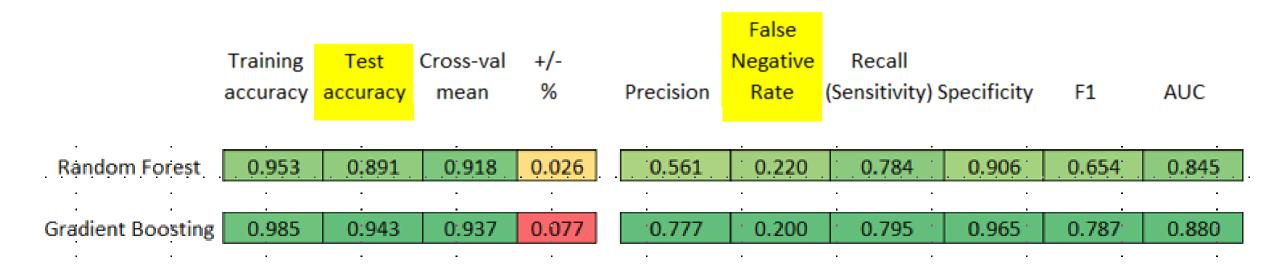
### Model Evaluation

	Training accuracy	Test accuracy	Cross-val mean	+/- %	Precision	False Negative Rate	Recall (Sensitivity)	Specificity	F1	AUC
Logistic Regression	0.803	0.745	0.793	0.011	0.308	0.250	0.750	0.744	0.437	0.747
Ridge Regression	0.800	0.738	0.792	0.011	0.301	0.250	0.750	0.736	0.430	0.743
Naive Bayes	0.795	0.738	0.787	0.057	0.233	0.570	0.432	0.784	0.303	0.608
KNN	0.994	0.783	0.896	0.004	0.206	0.770	0.227	0.867	0.216	0.547
Decision Tree	0.912	0.900	0.841	0.031	0.606	0.320	0.682	0.933	0.642	0.807
Random Forest	0.953	0.891	0.918	0.026	0.561	0.220	0.784	0.906	0.654	0.845
SVM	0.808	0.750	0.798	0.013	0.311	0.250	0.750	0.748	0.440	0.749
Gradient Boosting	0.985	0.943	0.937	0.077	0.777	0.200	0.795	0.965	0.787	0.880

Goal: Achieve high accuracy and reduce false negatives

– Don't want 'misses'

### Model Evaluation

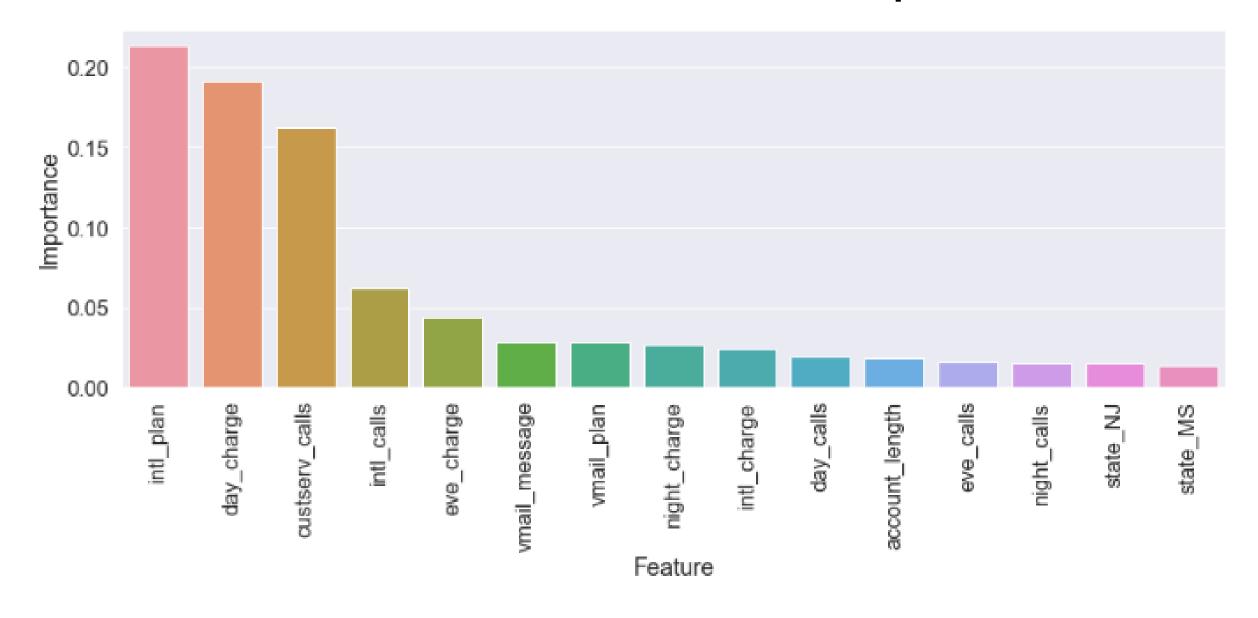


Gradient Boosting scored highest on most metrics.

Random Forest scored high and was the most consistent.



# Random Forest – Feature Importance



#### Business Initiatives to Reduce Churn

- Investigate possible operational issues International Plan
- Market Voice Mail Plan to increase loyalty
- Flag customer service call incidents for triage
- Use feature thresholds to trigger early action
- Break customer base into microsegments
- Use structured testing methodologies to evaluate combinations of value, messaging, and delivery method

### **Future Analysis**

#### Collect additional data on:

- Product types across departments
- Offers / Usage
- Call center data
- Network experience (Quality of Service)

# Q&A

