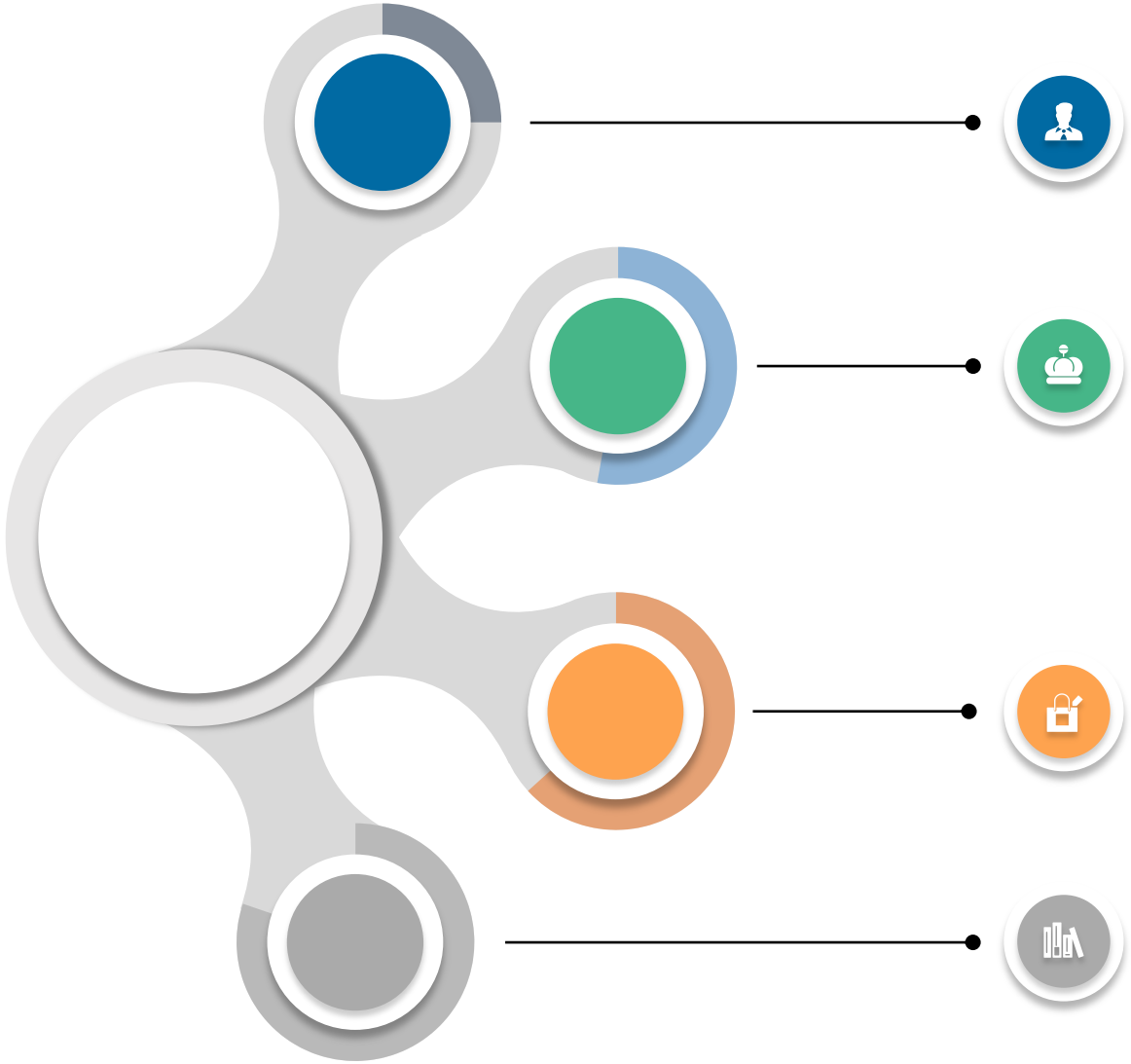
Predicting Churn in Telecommunications

Supervised Learning Capstone

Will Becerra

Agenda



Business Issue

The Data

Modeling

Takeaways

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

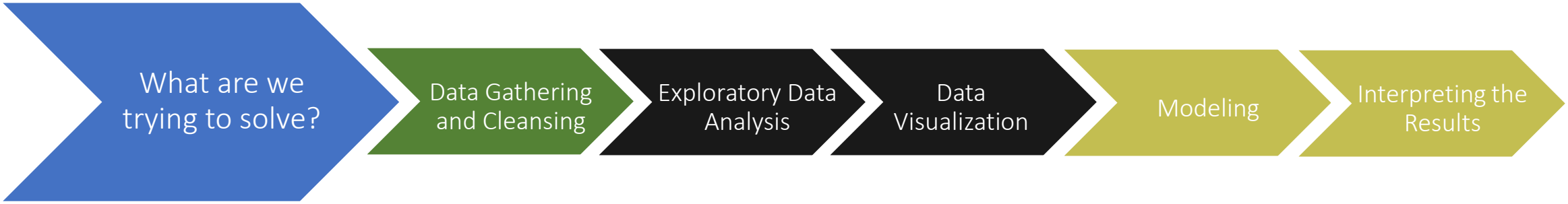
Disruption from OTT players

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

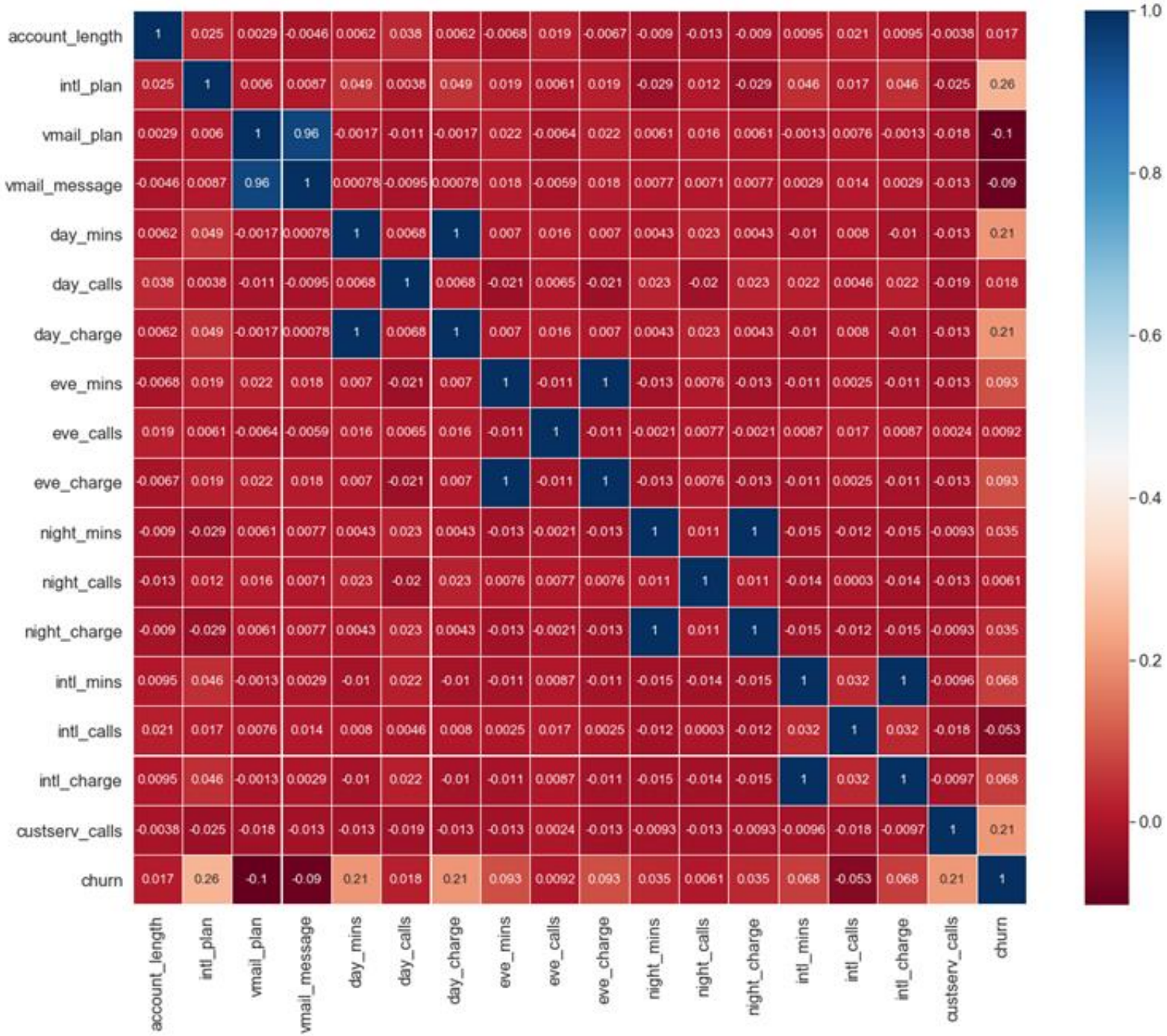
- State
- Area Code
- Phone
- Int'l Plan: yes/no
- VMail Plan: yes/no
- VMail Message: voice mails / per month
- Day Mins, Day Calls, Day Charge
- Eve Mins, Eve Calls, Eve Charge
- Night Mins, Night Calls, Night Charge
- Intl Mins, Intl Calls, Intl Charge
- CustServ Calls
- Account Length: tenure with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

Github data set - Telecommunications Churn https://github.com/albahnsen/ML_RiskManagement/blob/master/datasets/churn.csv

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.

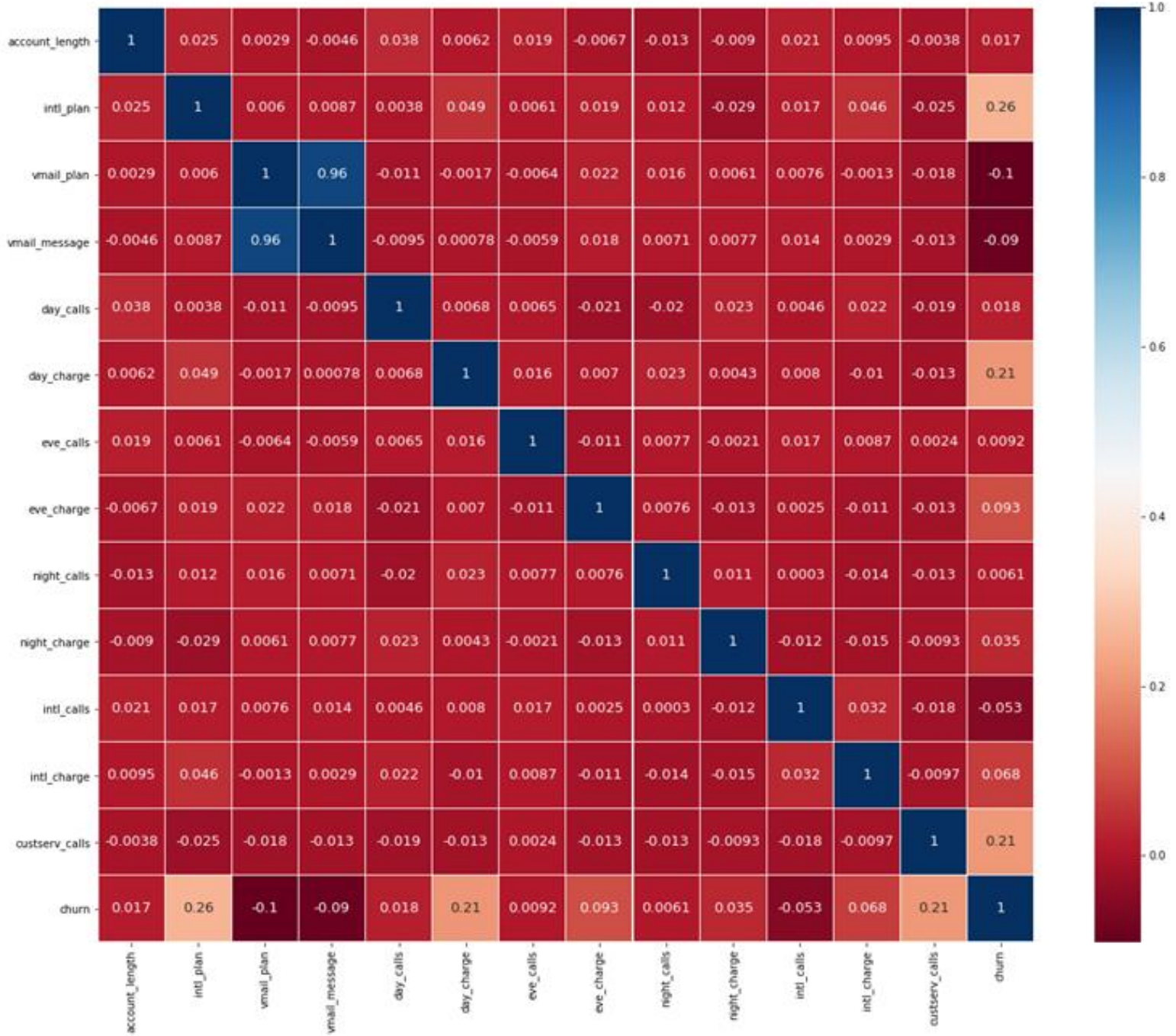
Exploratory Data Analysis



Pearson Correlation

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)



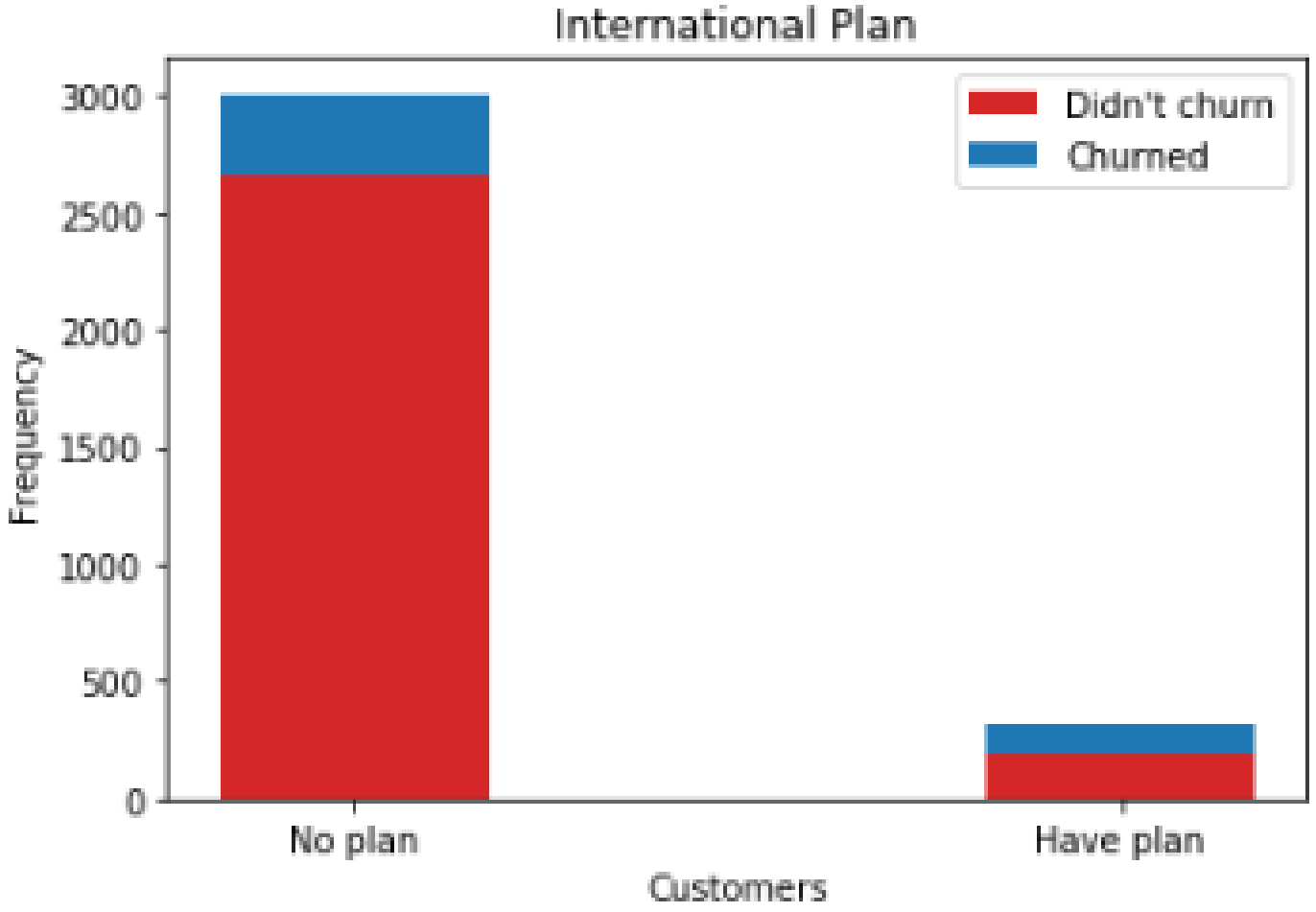
Correlation with churn

Intl Plan: 0.26

Day charge: 0.21

Customer service calls: 0.21

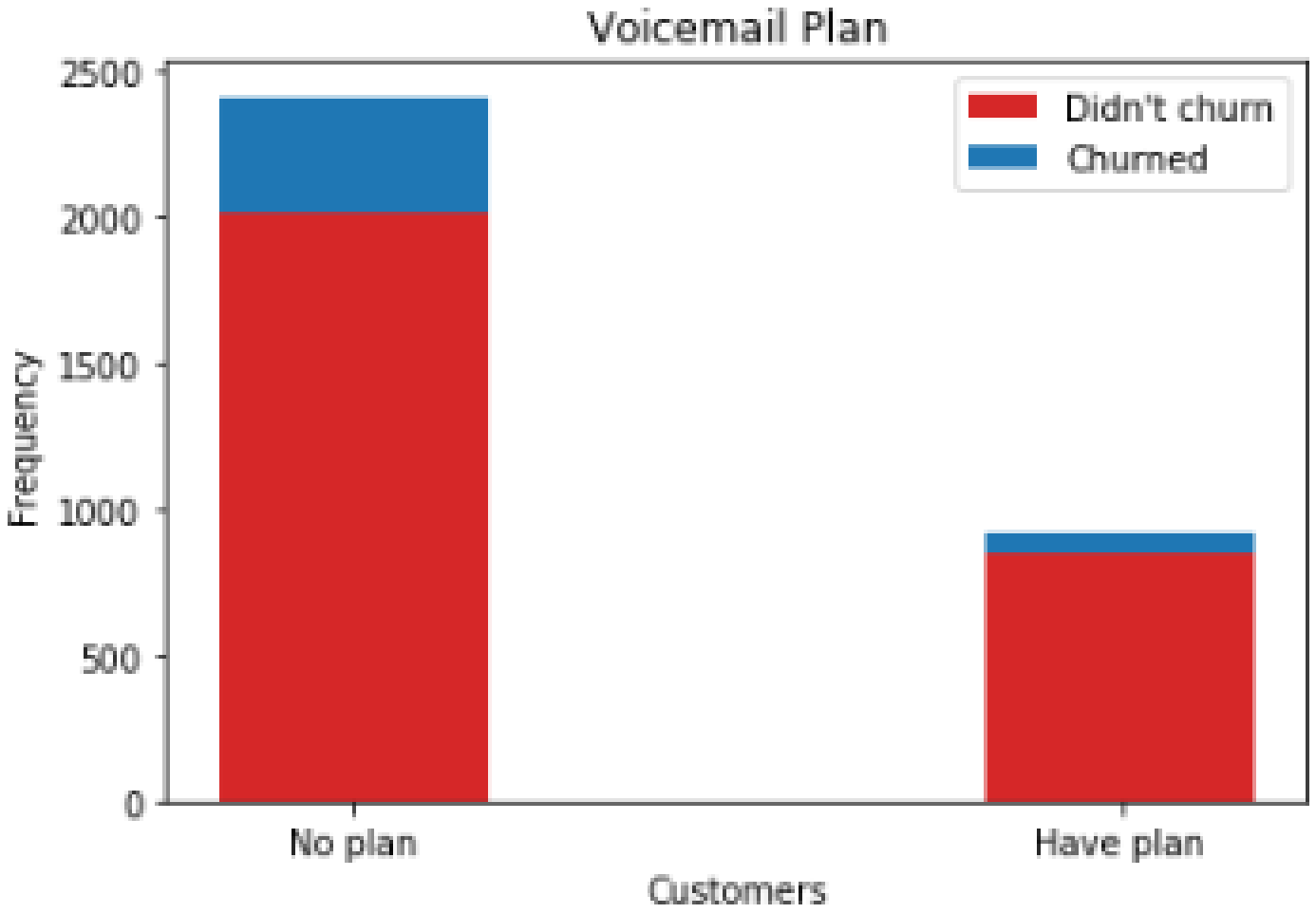
Customers with International Plan



Churn	No plan	Have plan	Total
True	346 *11.5%*	137 *42.4%*	483 *14.5%*
False	2664 *88.5%*	186 *57.6%*	2850 *85.5%*
Total	3010 *90.3%*	323 *9.7%*	3333

With plan, more than 3X as likely to leave

Customers with Voicemail Plan



Churn	No plan	Have plan	Total
True	403 16.7%	80 8.7%	483 14.5%
False	2008 83.3%	842 91.3%	2850 85.5%
Total	2411 72.3%	922 27.7%	3333

Without plan, more than 2X as likely to leave

Modeling

Divided data set train/test – 20%

Dealing with Imbalance of Classes

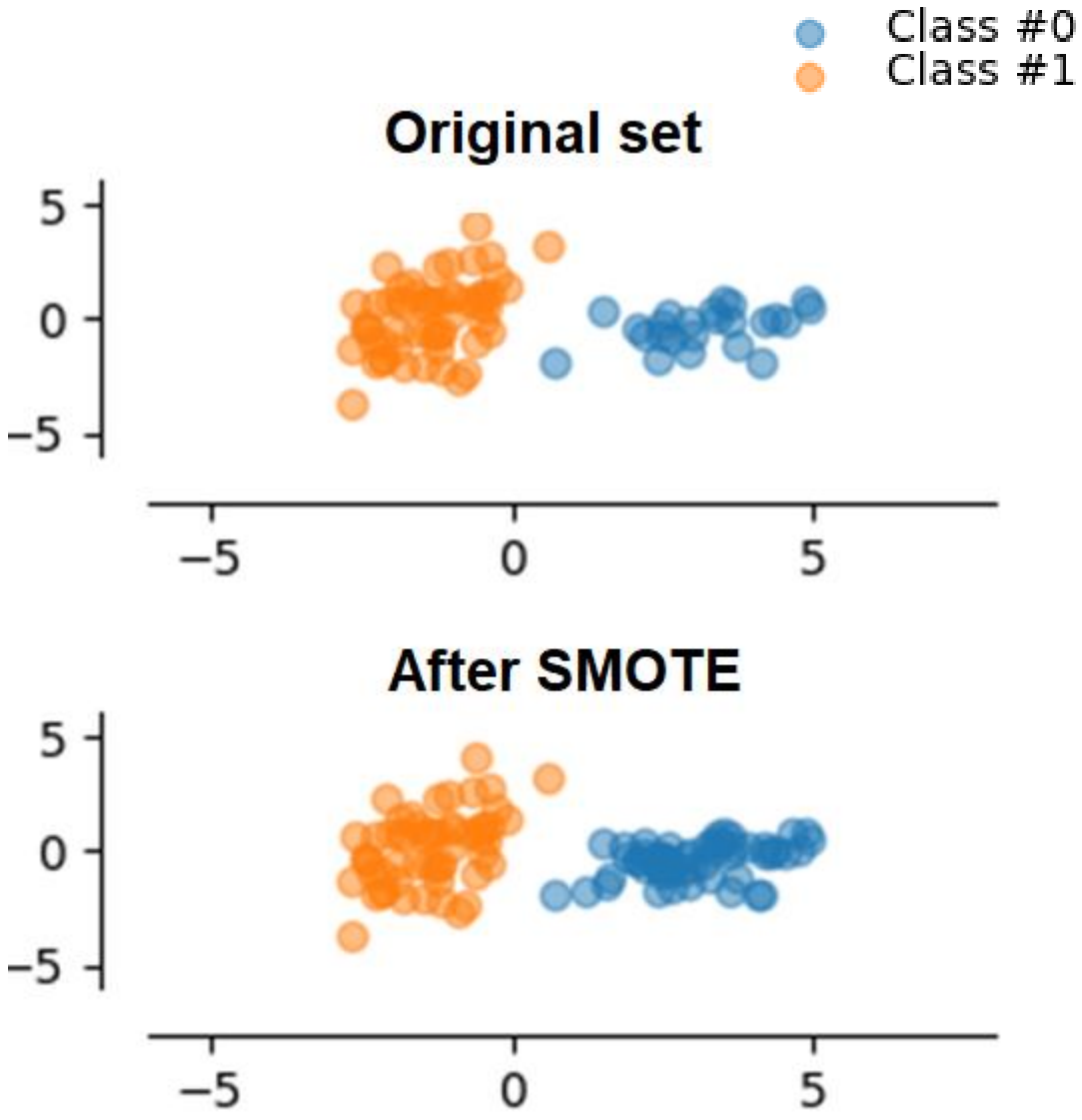
Only 14.9 % churned ➔ class imbalance

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

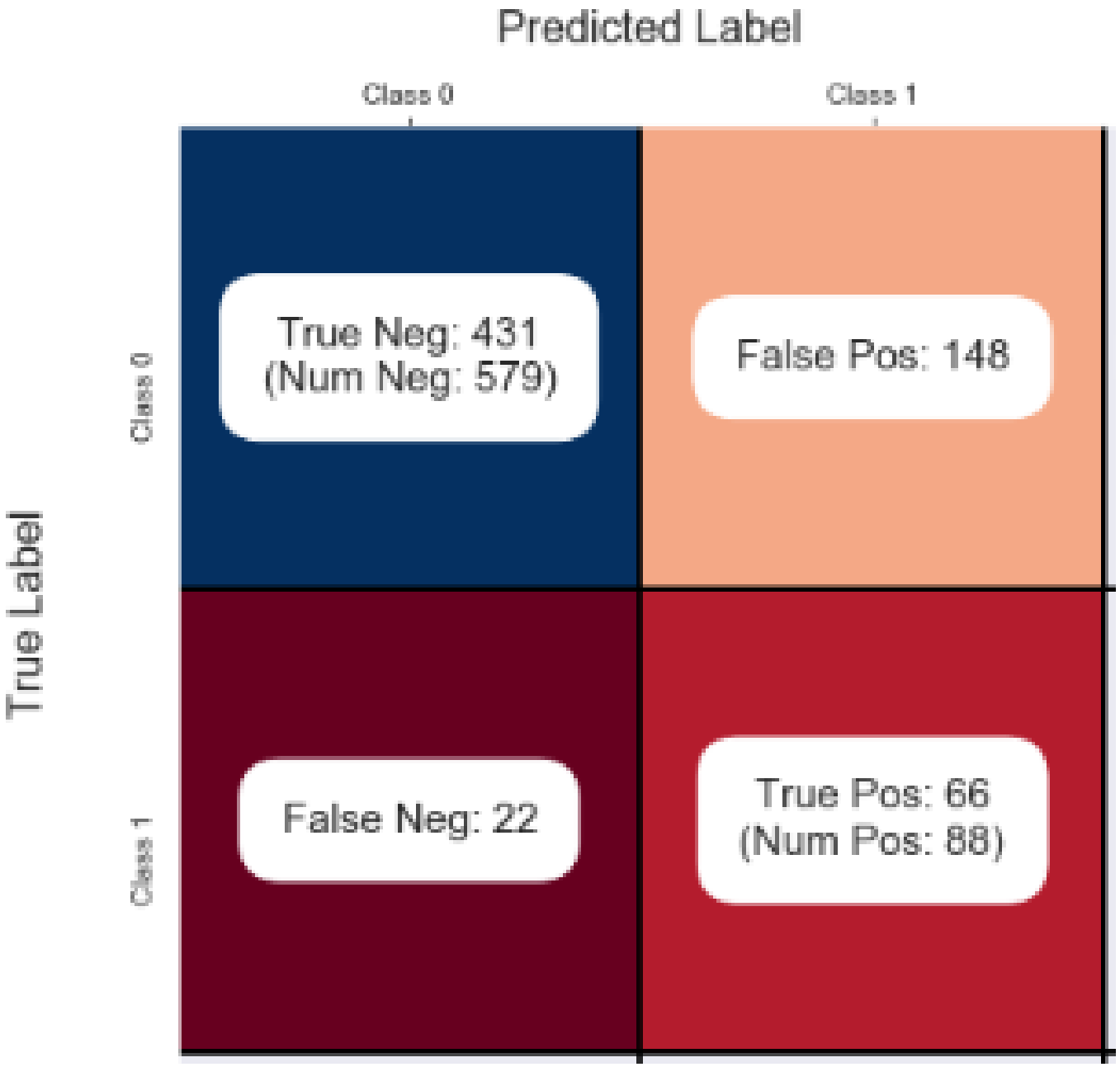
What if we guessed that no one churned?

Null Accuracy

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

```
0.8680659670164917
```

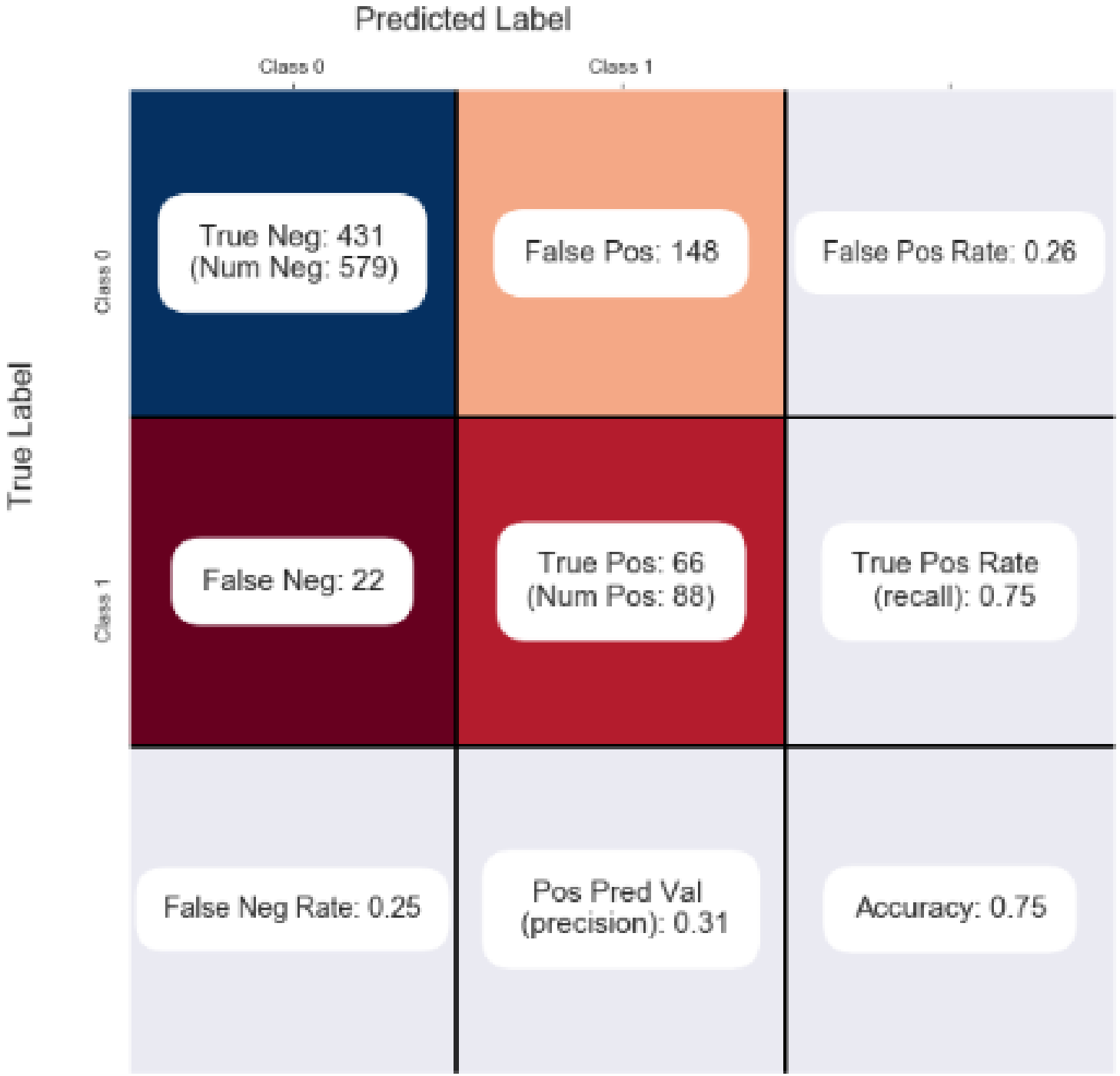
Guess right 86.8% of the time



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

(average false neg rate: 0.40)

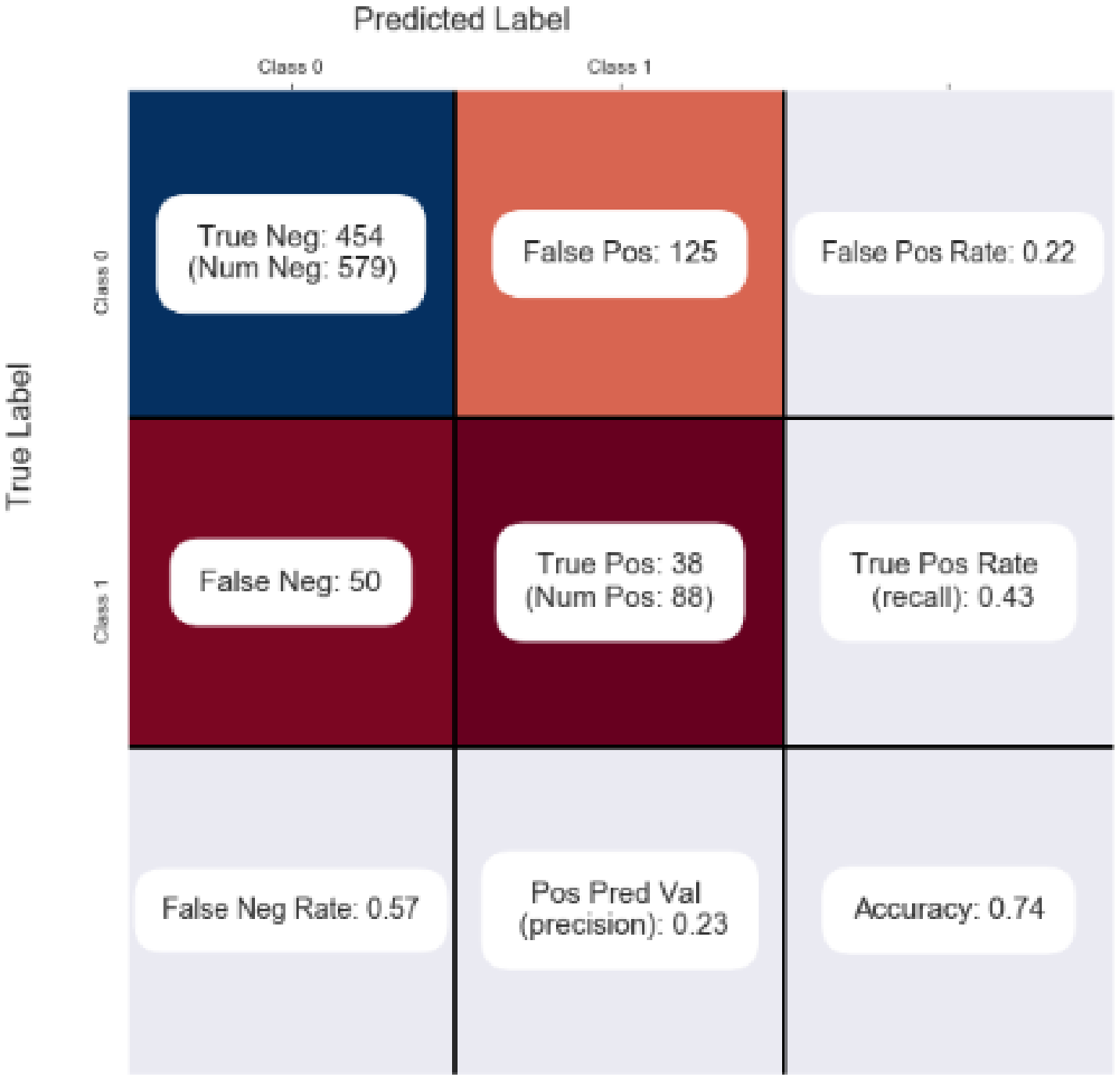
Ridge Classifier

Accuracy: 0.738

Cross-validation: 0.792 +/- 0.011

False negative rate: 0.25

(average false neg rate: 0.40)



Naïve Bayes

Accuracy: 0.738

Cross-validation: 0.787 +/- 0.057

False negative rate: 0.57

(average false neg rate: 0.40)

K Nearest Neighbors

Training accuracy: 0.994

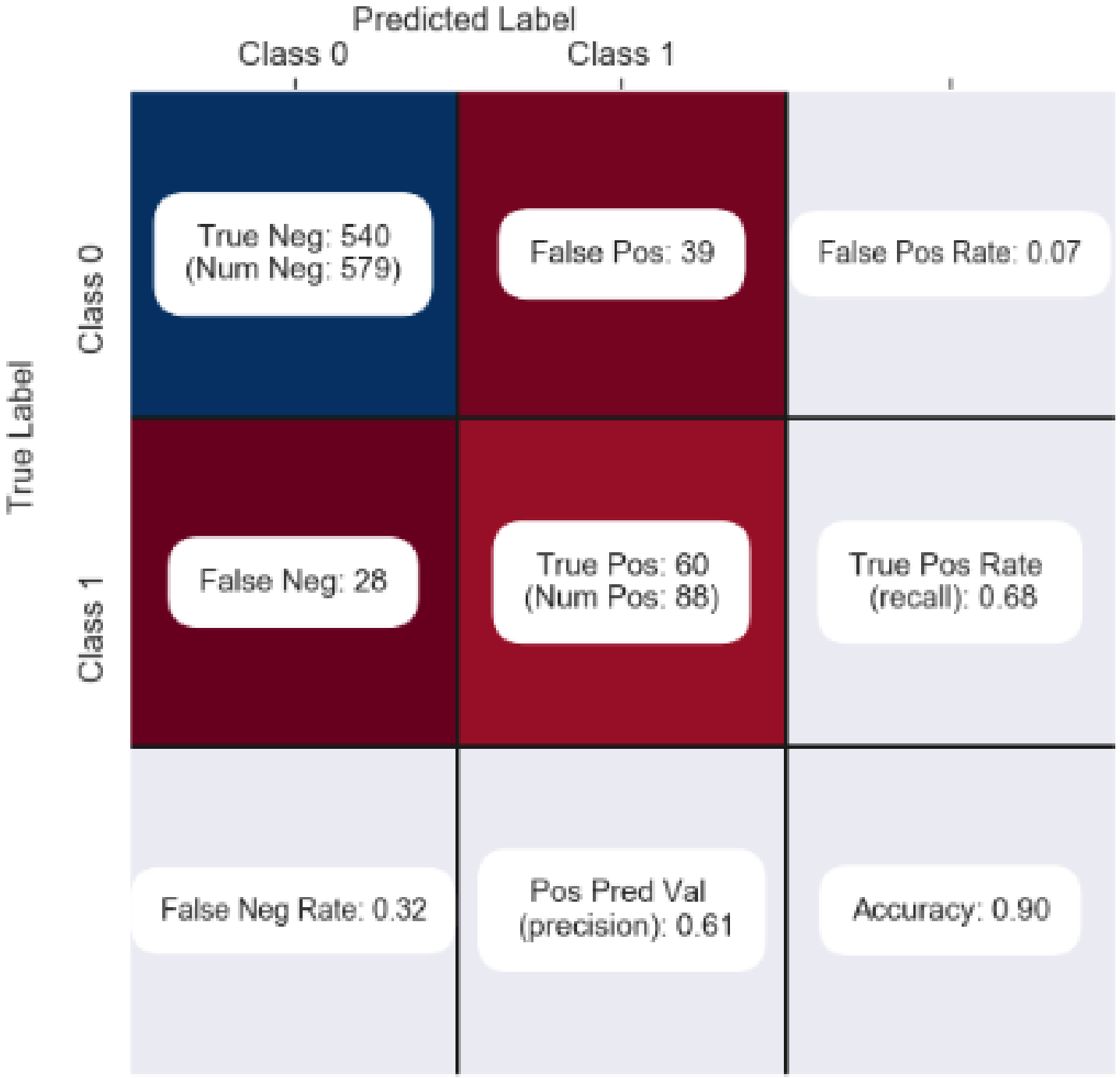
Testing accuracy: 0.783

Difference = 0.211

Seriously overfitting

False negative rate: 0.77

(average false neg rate: 0.40)



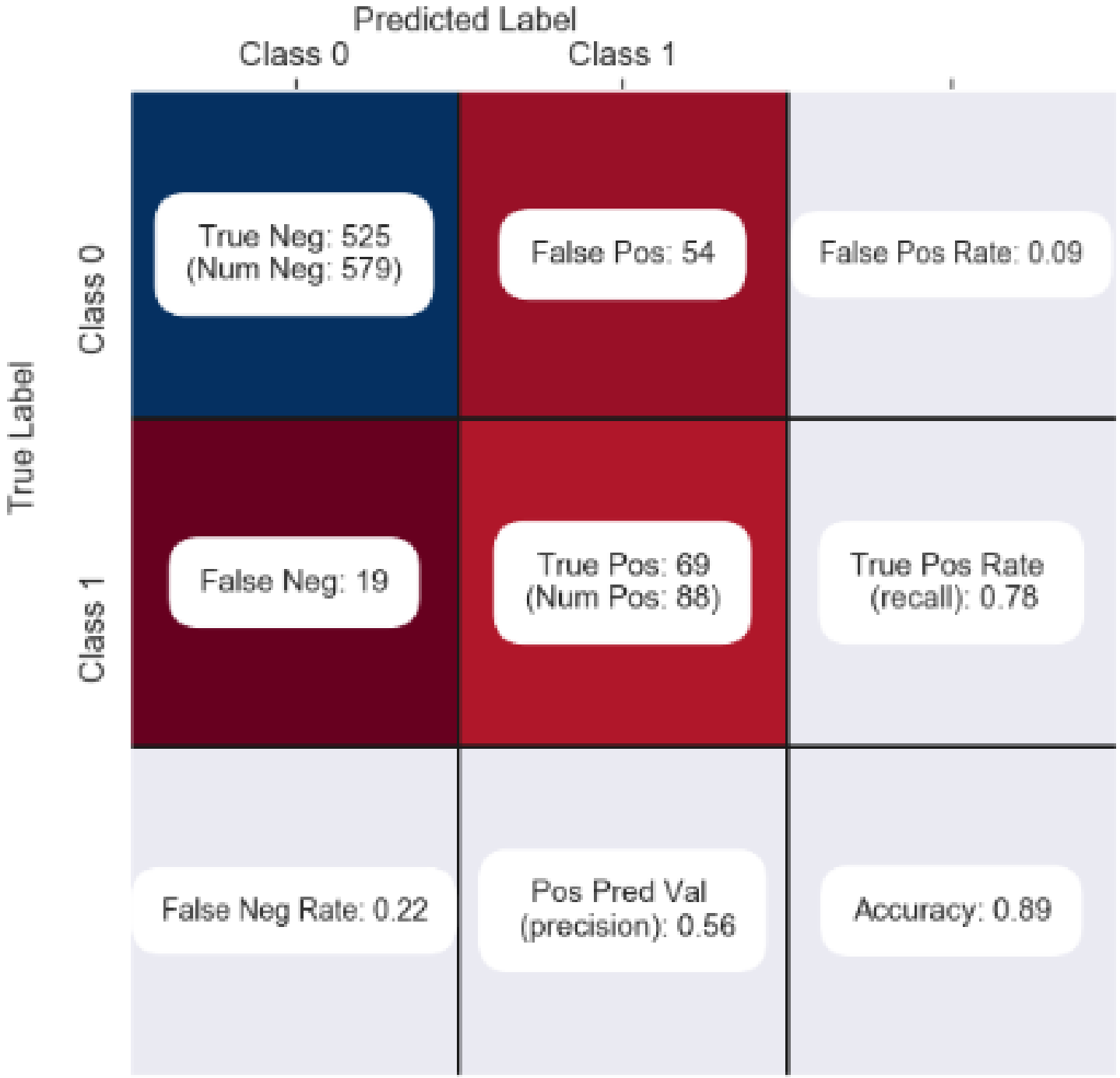
Decision Tree

Accuracy: 0.900

Cross-validation: 0.841 +/- 0.031

False negative rate: 0.32

(average false neg rate: 0.40)



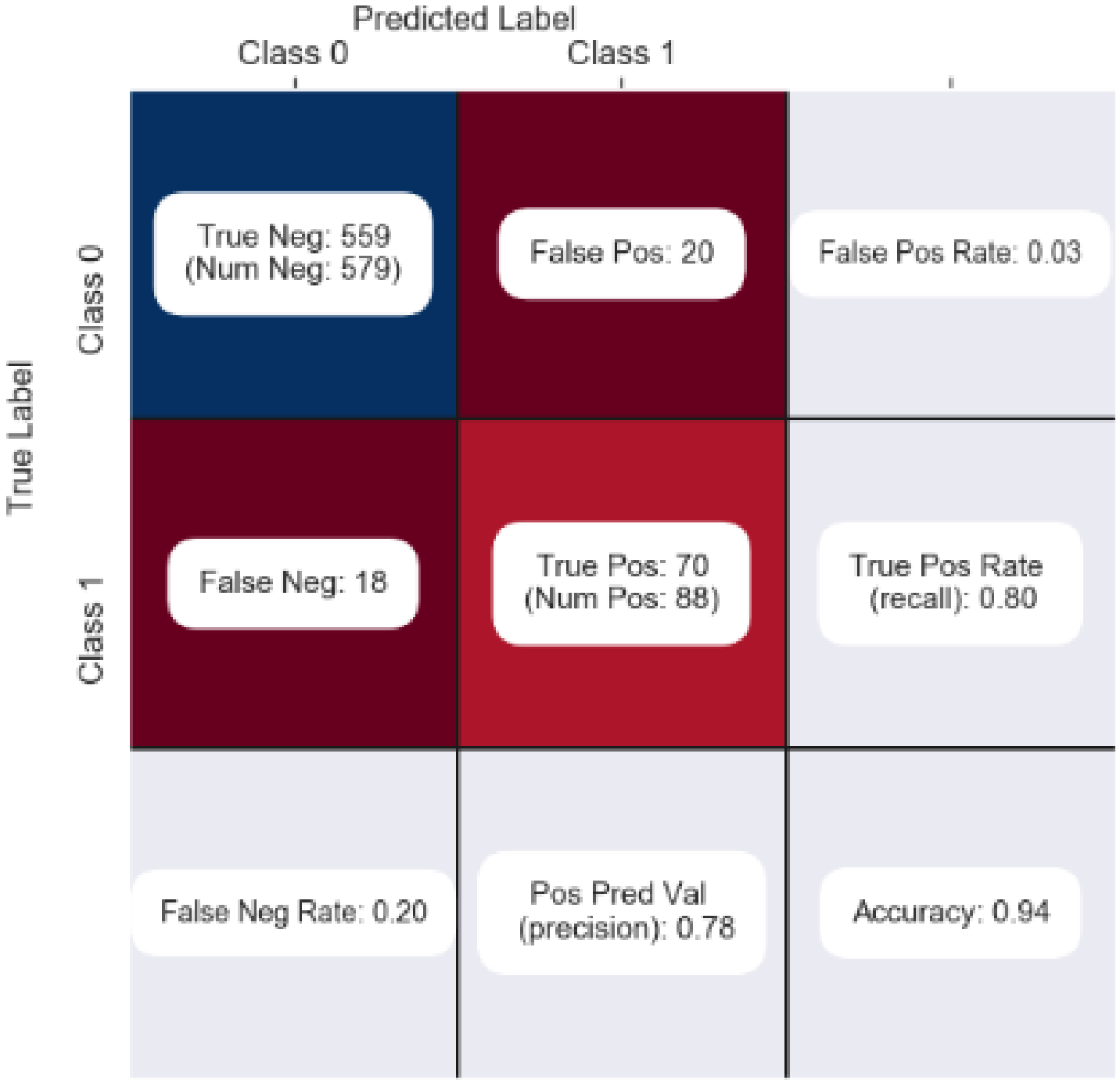
Random Forest

Accuracy: 0.891

Cross-validation: 0.918 +/- 0.026

False negative rate: 0.22

(average false neg rate: 0.40)



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’

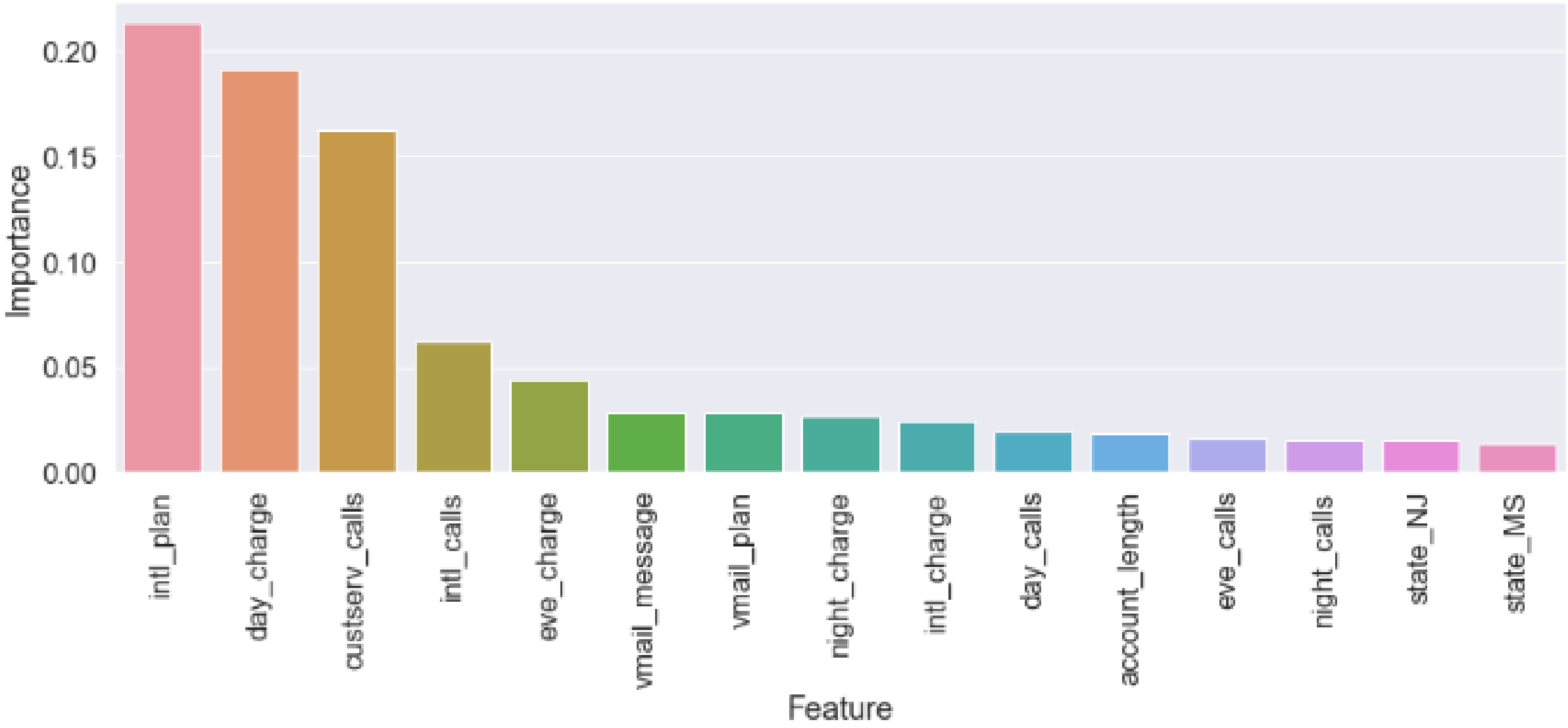
Model Evaluation

	Training accuracy	Test accuracy	Cross-val mean	+/- %	Precision	False Negative Rate	Recall (Sensitivity)	Specificity	F1	AUC
Random Forest	0.953	0.891	0.918	0.026	0.561	0.220	0.784	0.906	0.654	0.845
Gradient Boosting	0.985	0.943	0.937	0.077	0.777	0.200	0.795	0.965	0.787	0.880

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