KNOW YOUR CUSTOMER:

SYRIATEL CHURN MODEL

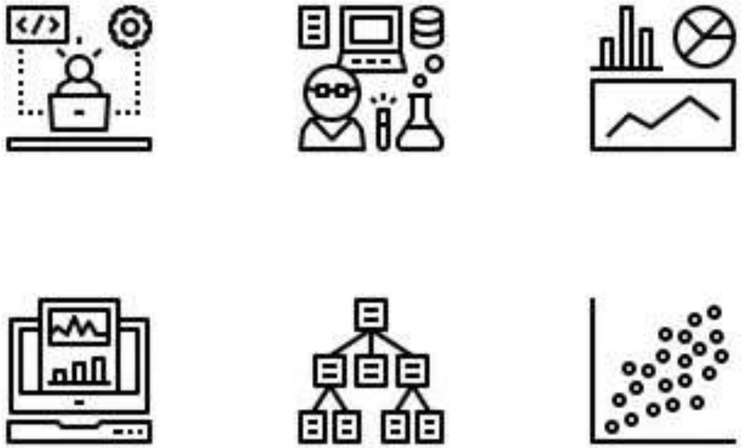
CONTENT

BUSINESS OBJECTIVE

Build a classifier to predict whether a customer will ("soon") stop doing business with SyriaTel, a telecommunications company.

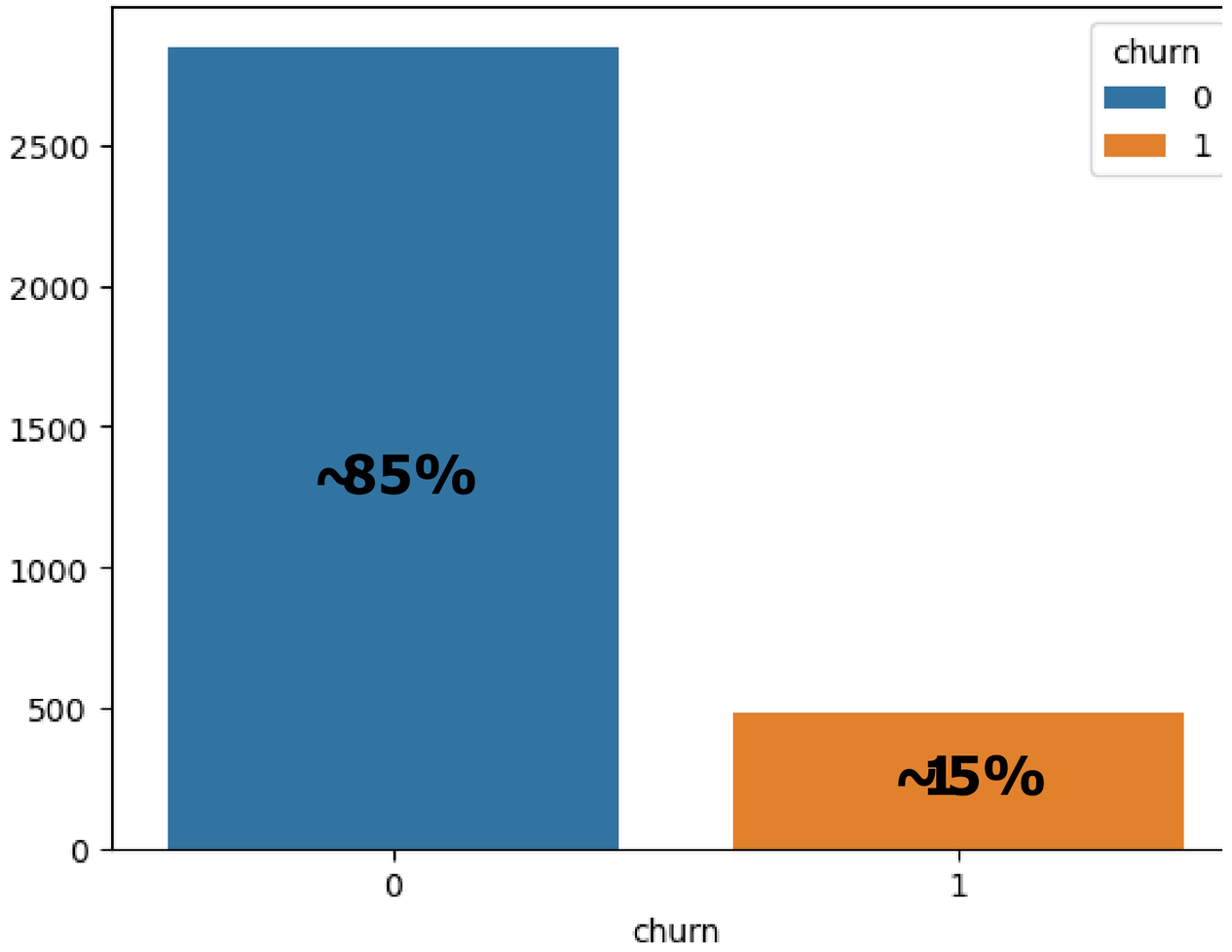
Measure Performance with Recall:

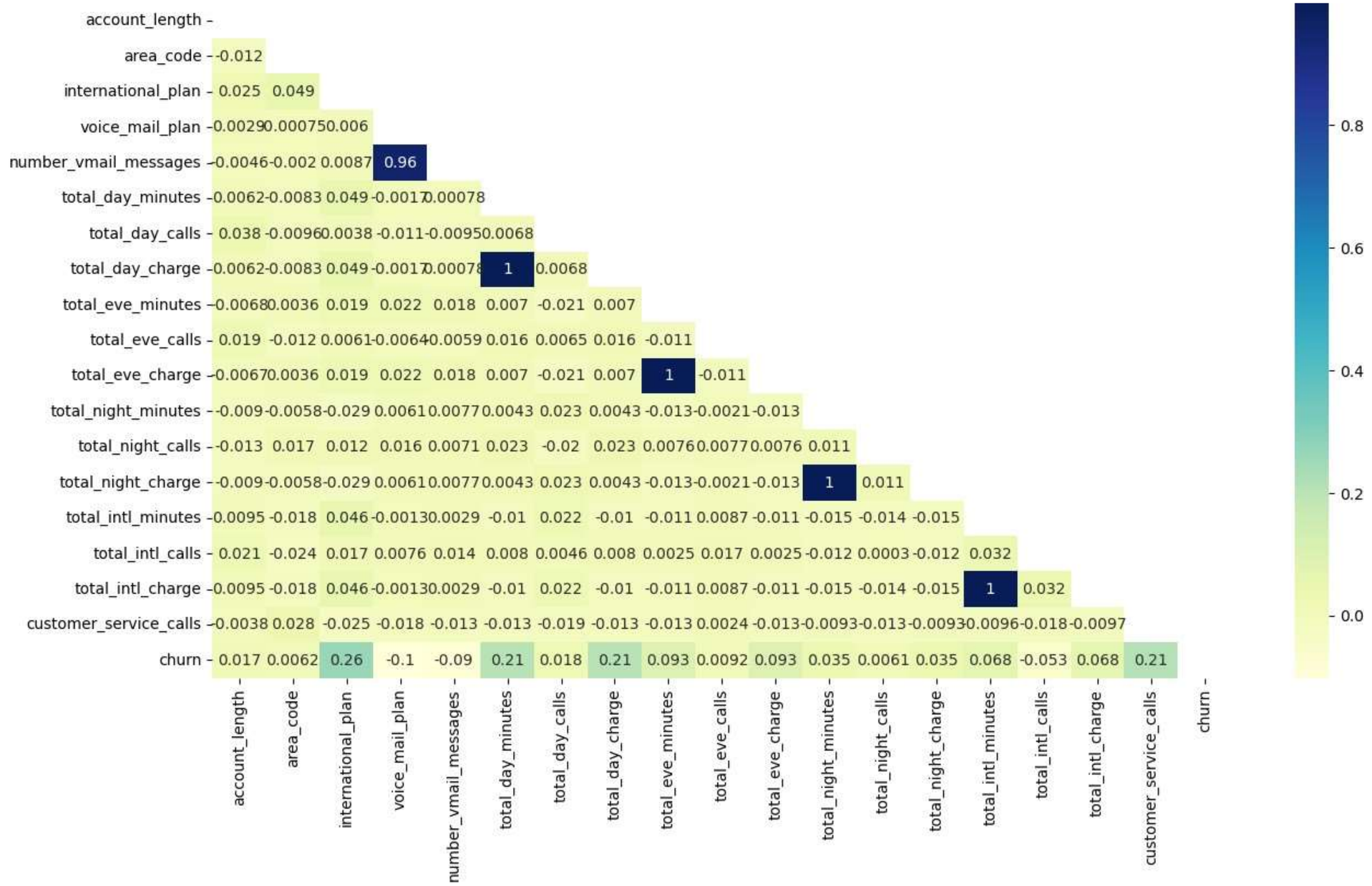
DATA ANALYSIS



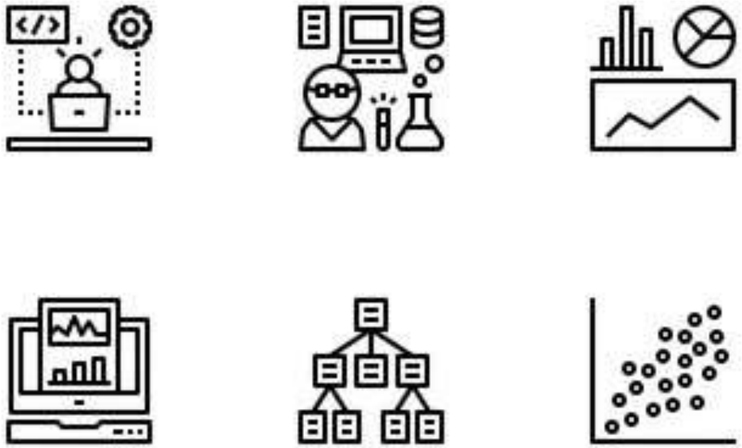
The Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3333 entries, 0 to 3332
Data columns (total 21 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   state                   3333 non-null   object 
 1   account length          3333 non-null   int64  
 2   area code               3333 non-null   int64  
 3   phone number            3333 non-null   object 
 4   international plan      3333 non-null   object 
 5   voice mail plan         3333 non-null   object 
 6   number vmail messages   3333 non-null   int64  
 7   total day minutes       3333 non-null   float64
 8   total day calls         3333 non-null   int64  
 9   total day charge        3333 non-null   float64
 10  total eve minutes       3333 non-null   float64
 11  total eve calls         3333 non-null   int64  
 12  total eve charge        3333 non-null   float64
 13  total night minutes     3333 non-null   float64
 14  total night calls       3333 non-null   int64  
 15  total night charge      3333 non-null   float64
 16  total intl minutes      3333 non-null   float64
 17  total intl calls        3333 non-null   int64  
 18  total intl charge       3333 non-null   float64
 19  customer service calls  3333 non-null   int64  
 20  churn                   3333 non-null   bool   
```




account_length
area_code -0.012
international_plan 0.025 0.049
voice_mail_plan -0.0029 0.00075 0.006
number_vmail_messages -0.0046 -0.002 0.0087 0.96
total_day_minutes -0.0062 -0.0083 0.049 -0.0017 0.00078
total_day_calls 0.038 -0.0096 0.0038 -0.011 -0.0095 0.0068
total_day_charge -0.0062 -0.0083 0.049 -0.0017 0.00078 1 0.0068
total_eve_minutes -0.0068 0.0036 0.019 0.022 0.018 0.007 -0.021 0.007
total_eve_calls 0.019 -0.012 0.0061 -0.0064 0.0059 0.016 0.0065 0.016 -0.011
total_eve_charge -0.0067 0.0036 0.019 0.022 0.018 0.007 -0.021 0.007 1 -0.011
total_night_minutes -0.009 -0.0058 -0.029 0.0061 0.0077 0.0043 0.023 0.0043 -0.013 -0.0021 -0.013
total_night_calls -0.013 0.017 0.012 0.016 0.0071 0.023 -0.02 0.023 0.0076 0.0077 0.0076 0.011
total_night_charge -0.009 -0.0058 -0.029 0.0061 0.0077 0.0043 0.023 0.0043 -0.013 -0.0021 -0.013 1 0.011
total_intl_minutes -0.0095 -0.018 0.046 -0.0013 0.0029 -0.01 0.022 -0.01 -0.011 0.0087 -0.011 -0.015 -0.014 -0.015
total_intl_calls 0.021 -0.024 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
total_intl_charge -0.0095 -0.018 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
customer_service_calls -0.0038 0.028 -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
churn 0.017 0.0062 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
account_length
area_code
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
0.8
0.6
0.4
0.2
0.0

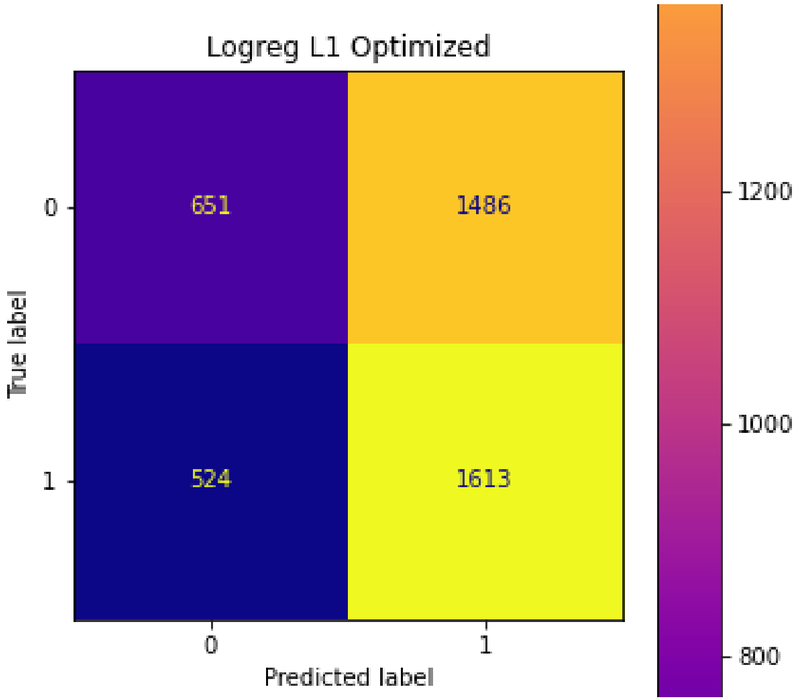
MODELING



Logistic Regression Models

Logistic L1 Model: Data Prep/Train Results

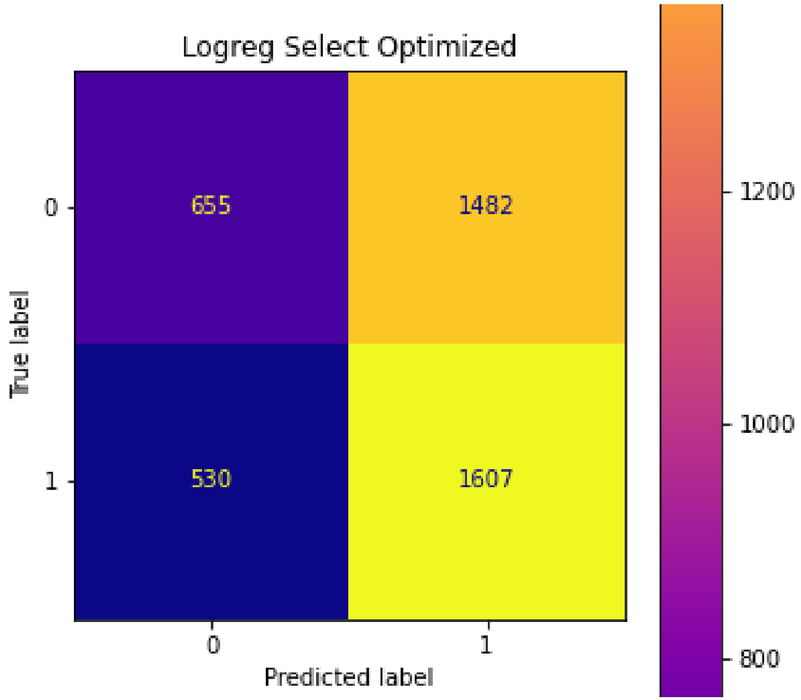
- This model contains **all predictor variables**, except phone number.



Processing steps
SMOTE, hyperparameter tuning, and OneHotEncoding

Logistic Select Model: Data Prep/Train Results

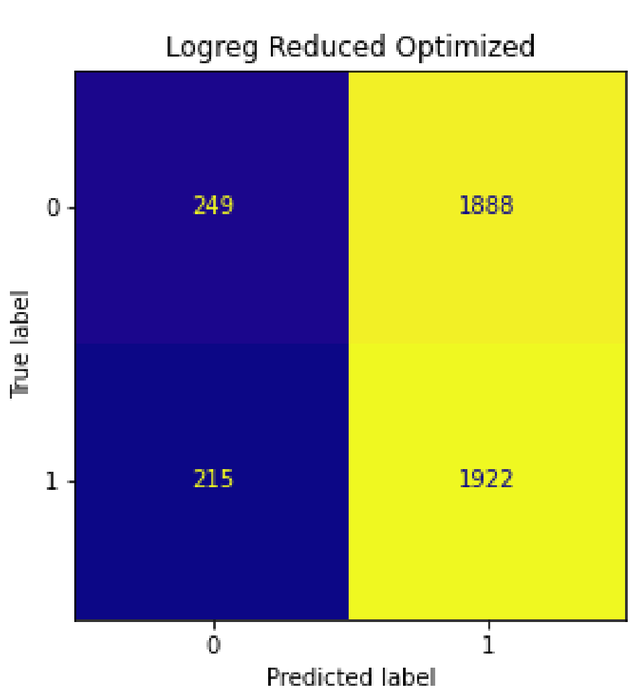
- We Used **SelectFromModel** to select **features** for us that are **most important**:
  - Reduced the dataframe from 69 predictors to 53.



Processing steps
SelectFromModel, hyperparameter tuning

Logistic Reduced Model: Data Prep/Train Results

- We only included highly correlated predictors with respect to churn.



Processing steps
SMOTE, hyperparameter tuning, and feature selection

Logistic Regression Models: Test Results

Train Results:

model_name	recall_score
Logistic L1	0.754796
Logistic Select	0.751989
Logistic Reduced	0.899392

Our Logistic Regression models are underfitting!!!!

Test Results:

Classification report for Model 1:

	precision	recall	f1-score	support
0	0.91	0.32	0.48	713
1	0.17	0.81	0.28	121
accuracy			0.39	834
macro avg	0.54	0.57	0.38	834
weighted avg	0.80	0.39	0.45	834

Classification report for Model 2:

	precision	recall	f1-score	support
0	0.91	0.33	0.48	713
1	0.17	0.81	0.28	121
accuracy			0.40	834
macro avg	0.54	0.57	0.38	834
weighted avg	0.80	0.40	0.45	834

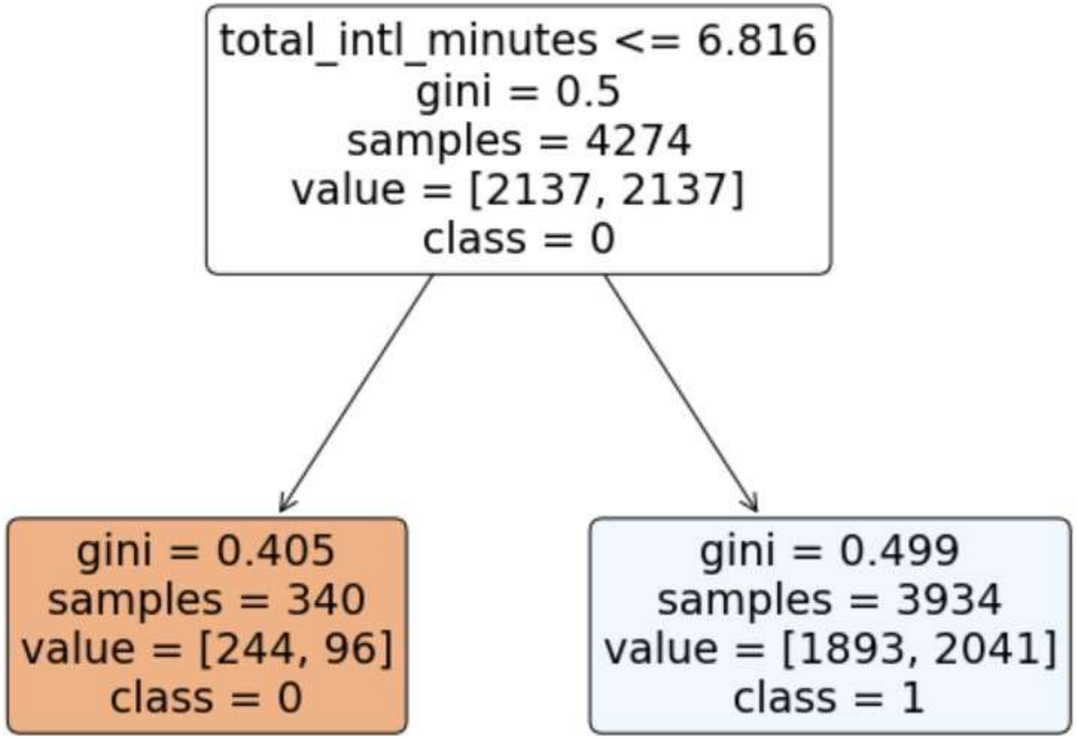
Classification report for Model 3:

	precision	recall	f1-score	support
0	0.96	0.13	0.24	713
1	0.16	0.97	0.27	121
accuracy			0.26	834
macro avg	0.56	0.55	0.25	834
weighted avg	0.84	0.26	0.24	834

DecisionTree Model

Data Prep/Train Results

- We used the same data that was used for model 3.
- Applied GridSearchCV() to find the optimal parameters for the model.
- Best Parameters:
  {'criterion': 'gini',
  'max_depth': 1,
  'max_features': 1,
  'min_samples_leaf': 1,
  'min_samples_split': 2,
  'splitter': 'best'}



CV Train Results

Metrics	Values
Mean Train Score	0.946065
Train Standard Deviation Score	0.022619
Mean Test Score	0.940583
Test Standard Deviation Score	0.027754

Test Results

	precision	recall	f1-score	support
0	0.89	0.12	0.21	713
1	0.15	0.92	0.26	121
accuracy			0.23	834
macro avg	0.52	0.52	0.23	834
weighted avg	0.78	0.23	0.21	834

FUTURE CONSIDERATIONS

Different type of model or ensemble modeling

STEP 1

STEP 2

Larger data set to offset the underfitting we encountered

Tiered Marketing Strategy

STEP 3