1 SyriaTel Customer Analysis

By: Brian Lee

1.1 Business problem

SyriaTel telecommunications company has hired us to analyze the causes to customer churn. **Churn** is whether a customer will stop doing business with the company. **The goal is to maintain SyriaTel customers in order to stably increase company profits and to build good customer reputation for future clientele.** We are trying to predict the likelihood of a customer churn based on a user's communication usage, plans, and other related factors.

Business Questions to have in mind:

- Do we need to improve our current customer plans?
- How does customer call usage affect their decisions to stay with SyriaTel?
- How is our response to customer needs?

1.2 Data Understanding

This project uses the SyriaTel customer dataset, which can be found in 'telecoms.csv' in the 'data' folder.

In [1]:
```
# Import necessary packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

import statsmodels.api as sm
import scipy.stats as stats
from sklearn.model_selection import train_test_split
from sklearn import preprocessing
from sklearn.preprocessing import OneHotEncoder

from sklearn.linear_model import LinearRegression
from sklearn import metrics

from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split, GridSearchCV

from sklearn.metrics import precision_score, recall_score, accuracy_score, f1_sco
from sklearn.metrics import confusion_matrix,recall_score,precision_recall_curve
from sklearn.metrics import precision_recall_fscore_support,f1_score,fbeta_score
from sklearn.metrics import classification_report, plot_roc_curve, plot_confusion
from sklearn.linear_model import LogisticRegression
from imblearn.over_sampling import SMOTE
from collections import Counter
from sklearn.metrics import classification_report

from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier

import xgboost as xgb
from xgboost import XGBClassifier
```

executed in 2.10s, finished 01:07:19 2021-04-25

In [2]:
```
df = pd.read_csv('data/telecoms.csv')
df.head()
```

executed in 52ms, finished 01:07:19 2021-04-25

Out[2]:

	state	account length	area code	phone number	international plan	voice mail plan	number vmail messages	total day minutes	total day calls	total day charge	...
0	KS	128	415	382-4657	no	yes	25	265.1	110	45.07	...
1	OH	107	415	371-7191	no	yes	26	161.6	123	27.47	...
2	NJ	137	415	358-1921	no	no	0	243.4	114	41.38	...
3	OH	84	408	375-9999	yes	no	0	299.4	71	50.90	...
4	OK	75	415	330-6626	yes	no	0	166.7	113	28.34	...

5 rows × 21 columns

1.3 Data Preparation

Let's quickly examine the dataset and clean it up for proper analysis and modeling

In [3]:
```
display(df.info())
display(df.describe())
```

executed in 75ms, finished 01:07:20 2021-04-25

```
<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   
dtypes: bool(1), float64(8), int64(8), object(4)
memory usage: 524.2+ KB
```

None

	account length	area code	number vmail messages	total day minutes	total day calls	total day charge	total eve minutes
count	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000
mean	101.064806	437.182418	8.099010	179.775098	100.435644	30.562307	200.980348
std	39.822106	42.371290	13.688365	54.467389	20.069084	9.259435	50.713844
min	1.000000	408.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	74.000000	408.000000	0.000000	143.700000	87.000000	24.430000	166.600000
50%	101.000000	415.000000	0.000000	179.400000	101.000000	30.500000	201.400000
75%	127.000000	510.000000	20.000000	216.400000	114.000000	36.790000	235.300000
max	243.000000	510.000000	51.000000	350.800000	165.000000	59.640000	363.700000

In [4]:
```
# Adjust column names with '_'
df.columns = df.columns.str.replace(' ','_')
df.info()
```

executed in 13ms, finished 01:07:20 2021-04-25

```
<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   
dtypes: bool(1), float64(8), int64(8), object(4)
memory usage: 524.2+ KB
```

In [5]:
```
# Find missing values columns
df.isna().sum()
```
executed in 13ms, finished 01:07:20 2021-04-25

Out[5]:
```
state                     0
account_length            0
area_code                 0
phone_number              0
international_plan        0
voice_mail_plan           0
number_vmail_messages     0
total_day_minutes         0
total_day_calls           0
total_day_charge          0
total_eve_minutes         0
total_eve_calls           0
total_eve_charge          0
total_night_minutes       0
total_night_calls         0
total_night_charge        0
total_intl_minutes        0
total_intl_calls          0
total_intl_charge         0
customer_service_calls    0
churn                     0
dtype: int64
```

- No missing values found.

Phone numbers are just unique identifiers. Does not provide additional information. Could drop it

In [6]:
```
# Drop phone_number
df.drop('phone_number', axis=1, inplace=True)
```
executed in 15ms, finished 01:07:20 2021-04-25

In [7]:
```
# Values for area_code
df.area_code.value_counts()
```
executed in 13ms, finished 01:07:20 2021-04-25

Out[7]:
```
415    1655
510     840
408     838
Name: area_code, dtype: int64
```

Taking a look at the area_code values show that there are only 3 area codes (San Francisco area), despite the data being declared for several different states. It will be better to drop the area codes in this case.

In [8]:
```
# Drop area_code column
df.drop('area_code', axis = 1, inplace=True)
```
executed in 13ms, finished 01:07:20 2021-04-25

Let's change the categorical columns to integers for easier analysis

In [9]:
```
# Change churn and plans to int
df['churn'] *= 1

df['international_plan'] = df['international_plan'].apply(lambda x: 1 if x=='yes'
df['voice_mail_plan'] = df['voice_mail_plan'].apply(lambda x: 1 if x=='yes' else

df[['churn', 'international_plan', 'voice_mail_plan']].astype(int)
```
executed in 29ms, finished 01:07:20 2021-04-25

Out[9]:

	churn	international_plan	voice_mail_plan
0	0	0	1
1	0	0	1
2	0	0	0
3	0	1	0
4	0	1	0
...	...	...	...
3328	0	0	1
3329	0	0	0
3330	0	0	0
3331	0	1	0
3332	0	0	1

3333 rows × 3 columns

In [10]:
```
df.head()
```
executed in 29ms, finished 01:07:20 2021-04-25

Out[10]:

	state	account_length	international_plan	voice_mail_plan	number_vmail_messages	total_day_mi
0	KS	128	0	1	25	
1	OH	107	0	1	26	
2	NJ	137	0	0	0	
3	OH	84	1	0	0	
4	OK	75	1	0	0	

1.4 Feature Engineering

Let's make a few features that will be useful to examine on a customer's overall phone usage.

In [11]:
```
# Creating a column for total customer charges by adding all the individual charg
df['total_customer_charges'] = df['total_day_charge'] + df['total_eve_charge'] +
df['total_customer_charges'].head()
```
executed in 14ms, finished 01:07:20 2021-04-25

Out[11]:
```
0    75.56
1    59.24
2    62.29
3    66.80
4    52.09
Name: total_customer_charges, dtype: float64
```

In [12]:
```
# Doing the same method for total calls and minutes
df['total_customer_calls'] = df['total_day_calls'] + df['total_eve_calls'] + df['
df['total_customer_minutes'] = df['total_day_minutes'] + df['total_eve_minutes']
df.head()
```
executed in 45ms, finished 01:07:20 2021-04-25

Out[12]:

	state	account_length	international_plan	voice_mail_plan	number_vmail_messages	total_day_mi
0	KS	128	0	1	25	
1	OH	107	0	1	26	
2	NJ	137	0	0	0	
3	OH	84	1	0	0	
4	OK	75	1	0	0	

5 rows × 22 columns

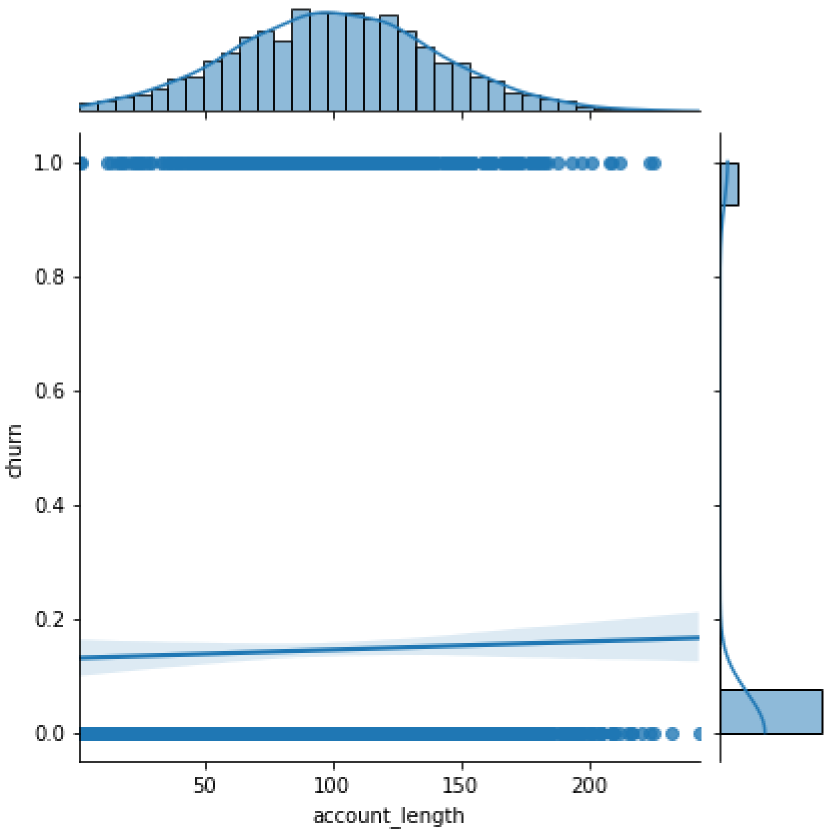
1.5 Exploratory Data Analysis

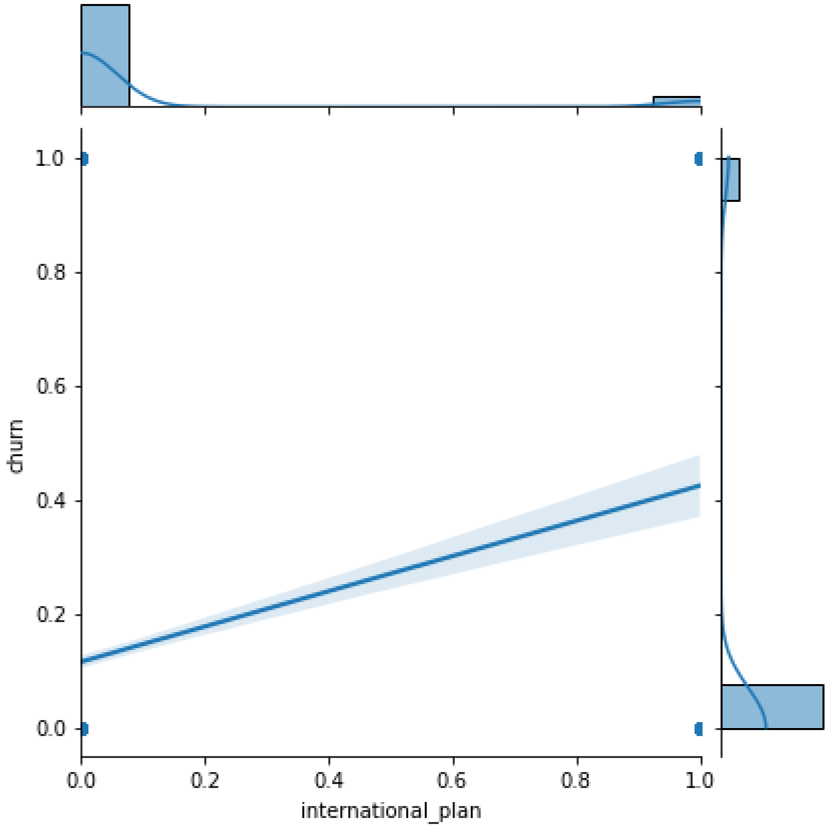
Using the cleaned data, we will examine the distributions of the columns and descriptive statistics for the dataset

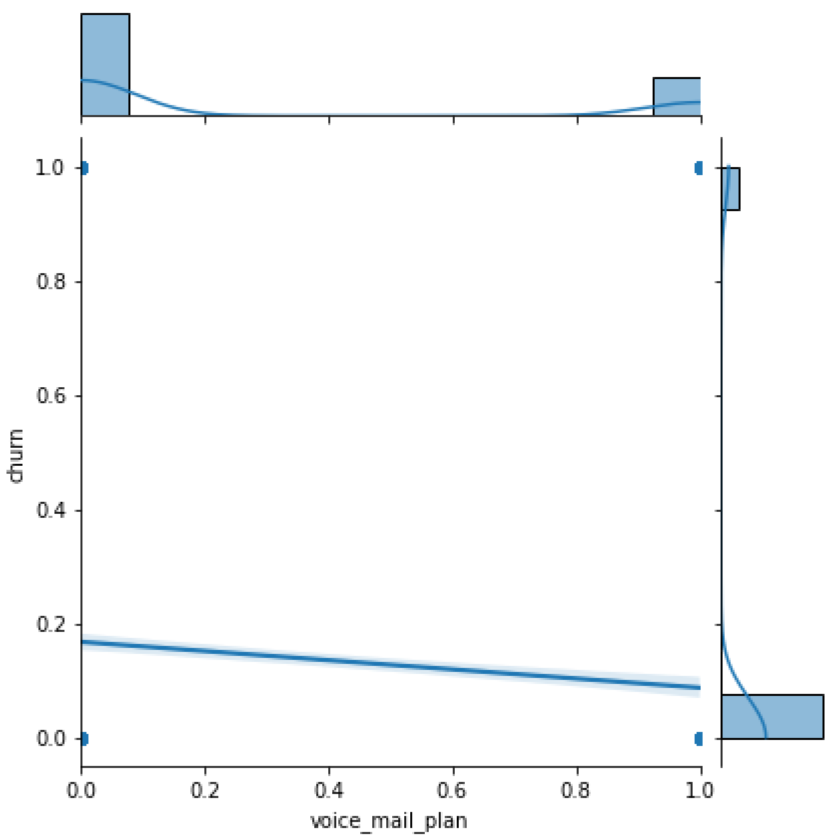
In [13]:
```
# Let's take a look at each column in relation with churn
for col in df.columns[1:]:
    sns.jointplot(x=col, y='churn', data=df, kind='reg');
```

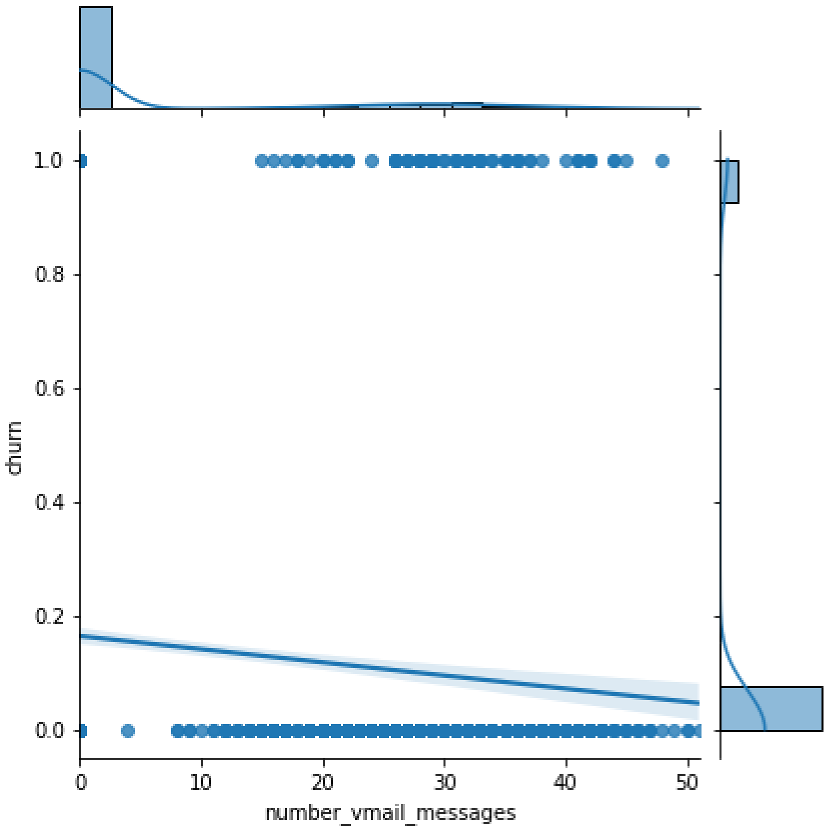
executed in 14.3s, finished 01:07:34 2021-04-25

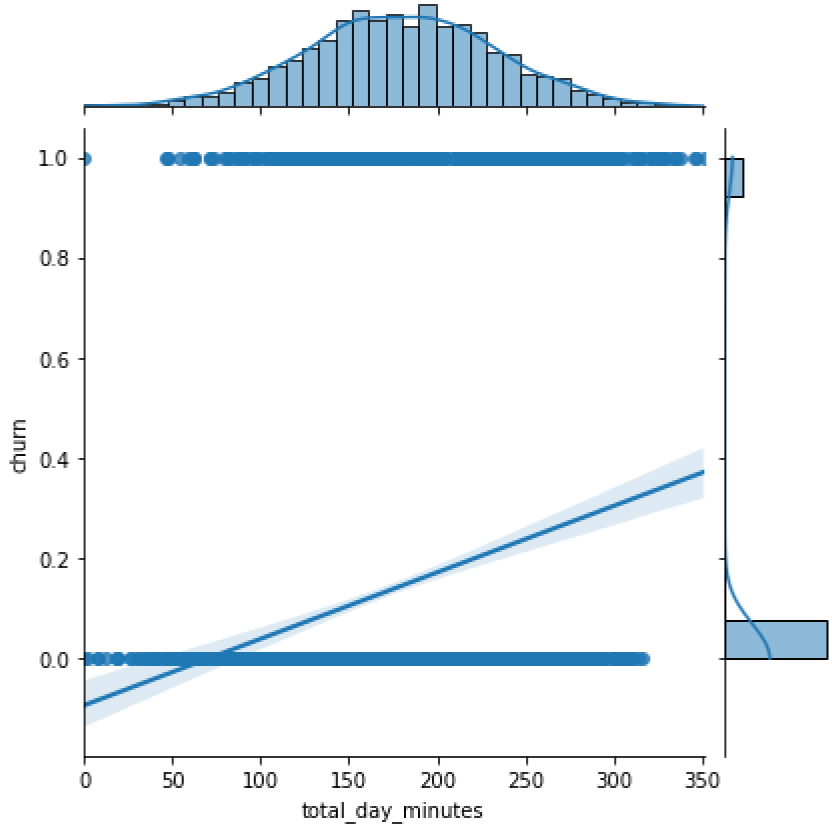
```
C:\Users\leebr\anaconda3\envs\learn-env\lib\site-packages\seaborn\axisgrid.py:1
559: RuntimeWarning: More than 20 figures have been opened. Figures created thr
ough the pyplot interface (`matplotlib.pyplot.figure`) are retained until expli
citly closed and may consume too much memory. (To control this warning, see the
rcParam `figure.max_open_warning`).
  f = plt.figure(figsize=(height, height))
```

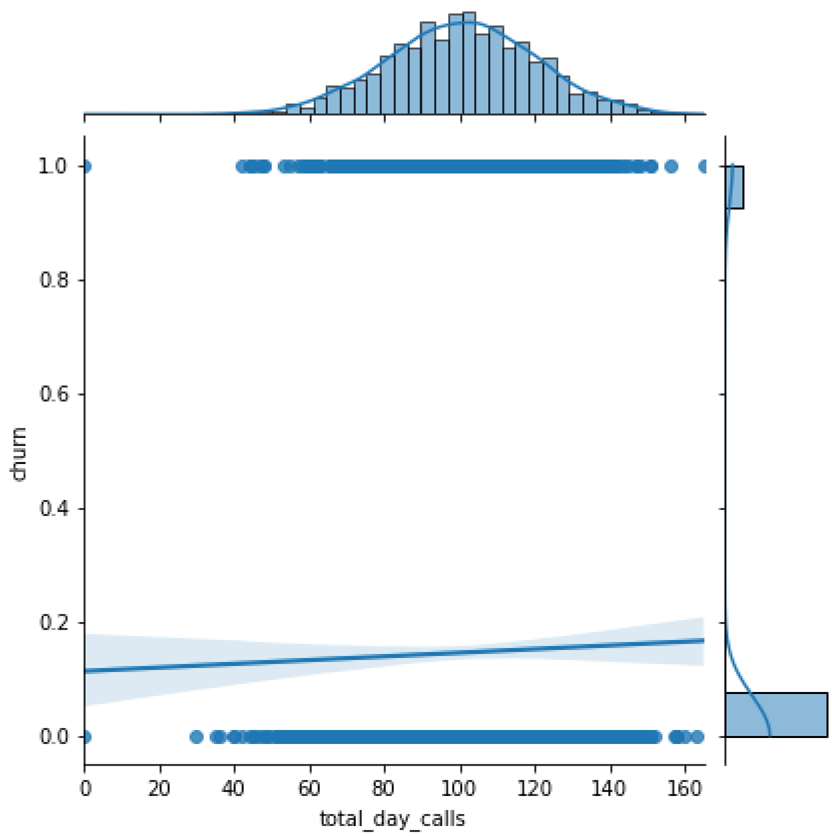


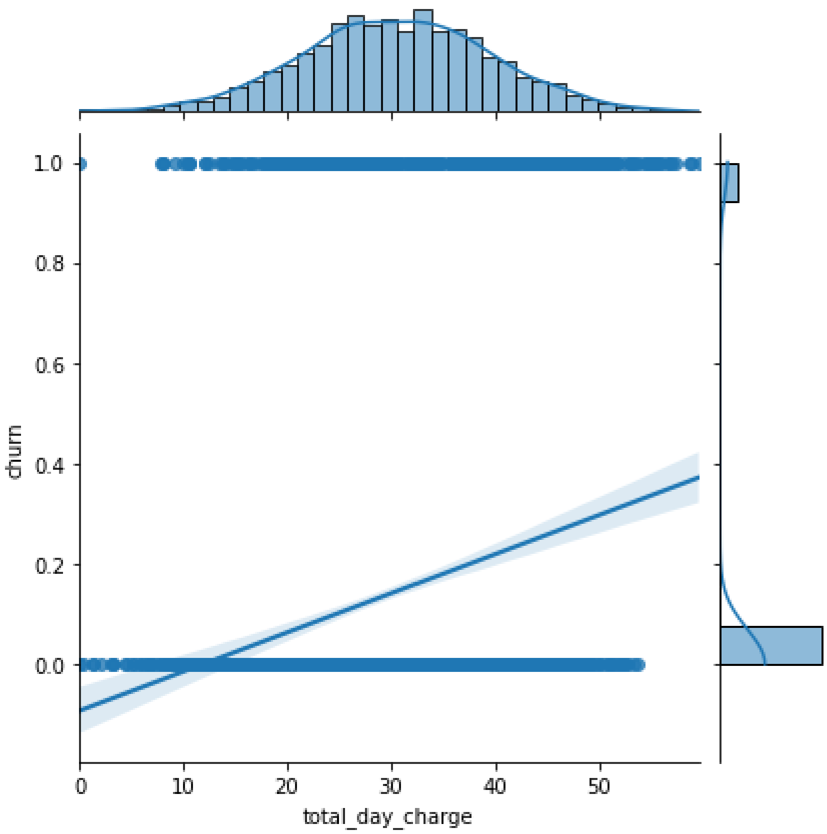

churn
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
international_plan

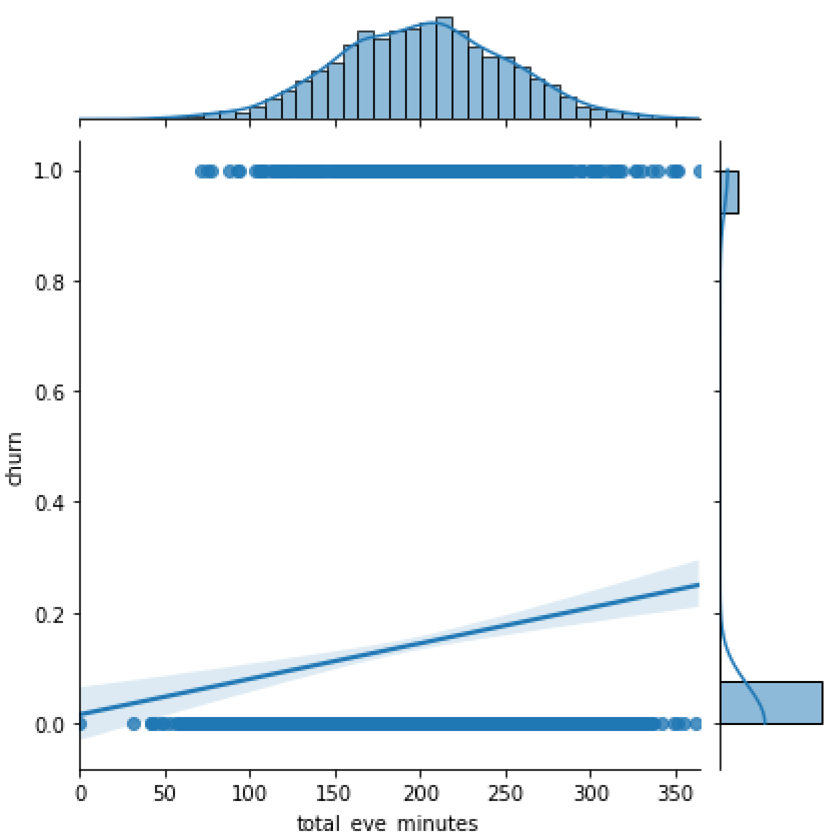

churn
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
voice_mail_plan

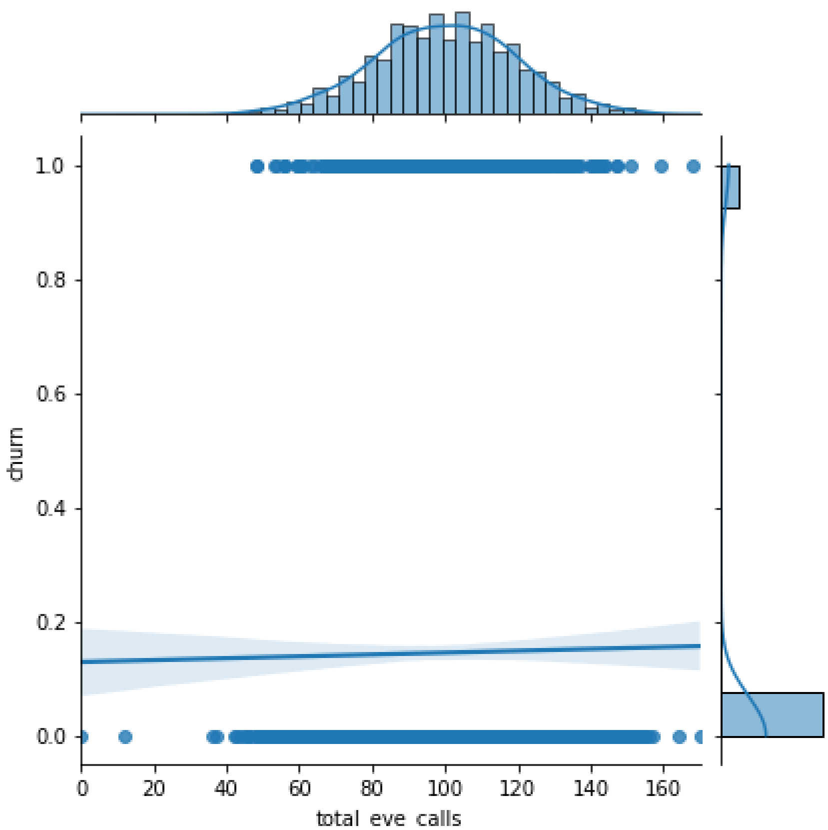

1.0
0.8
0.6
0.4
0.2
0.0
churn
0
10
20
30
40
50
number_vmail_messages

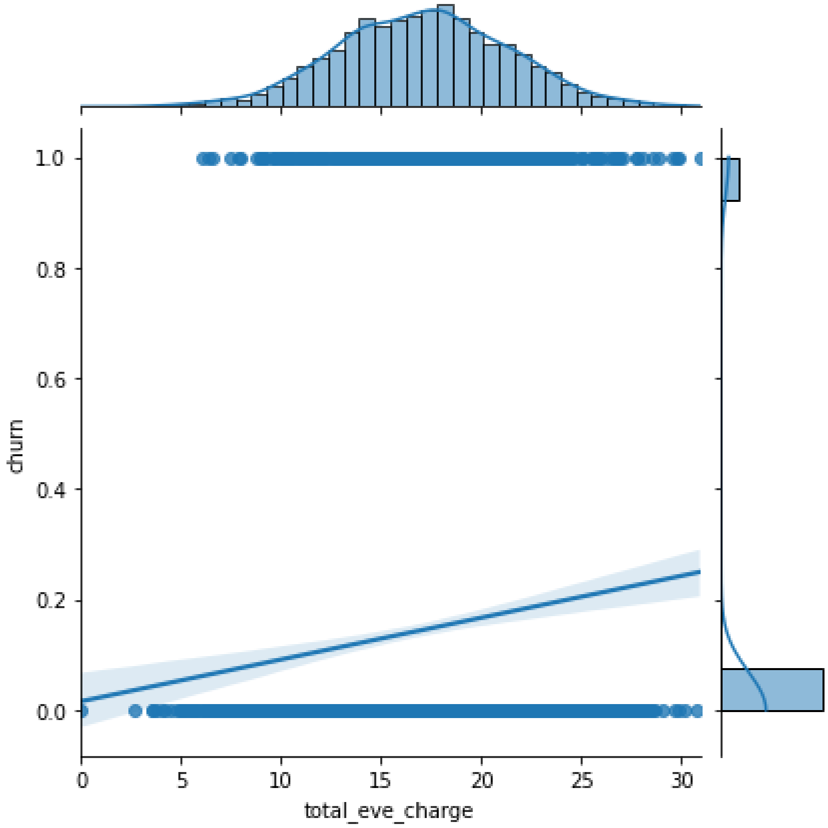

1.0
0.8
0.6
0.4
0.2
0.0
churn
0
50
100
150
200
250
300
350
total_day_minutes

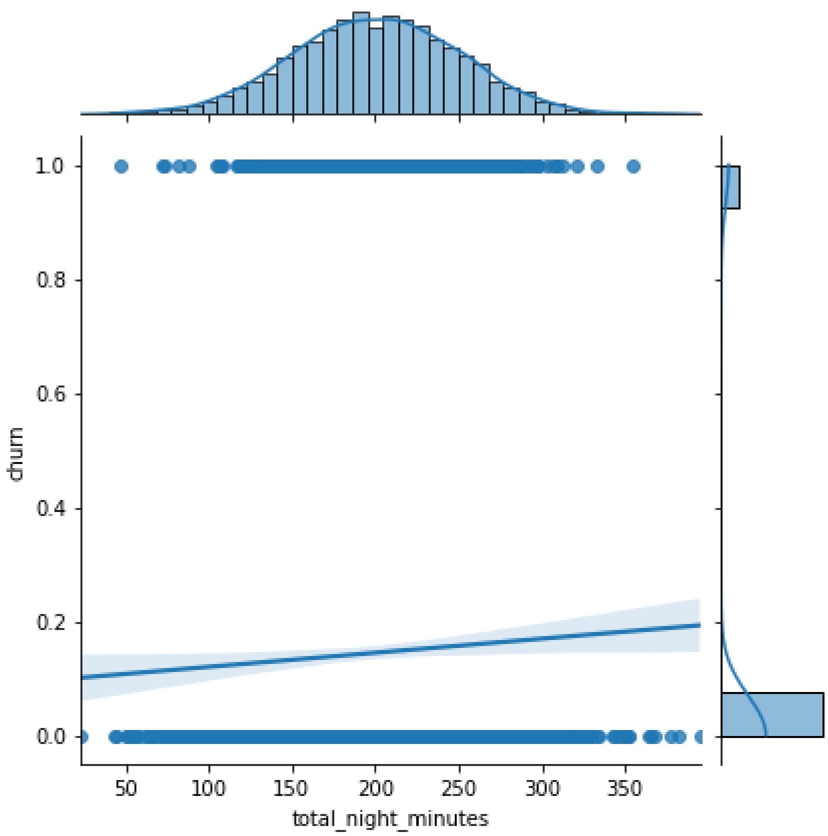

1.0
0.8
0.6
0.4
0.2
0.0
churn
0
20
40
60
80
100
120
140
160
total_day_calls

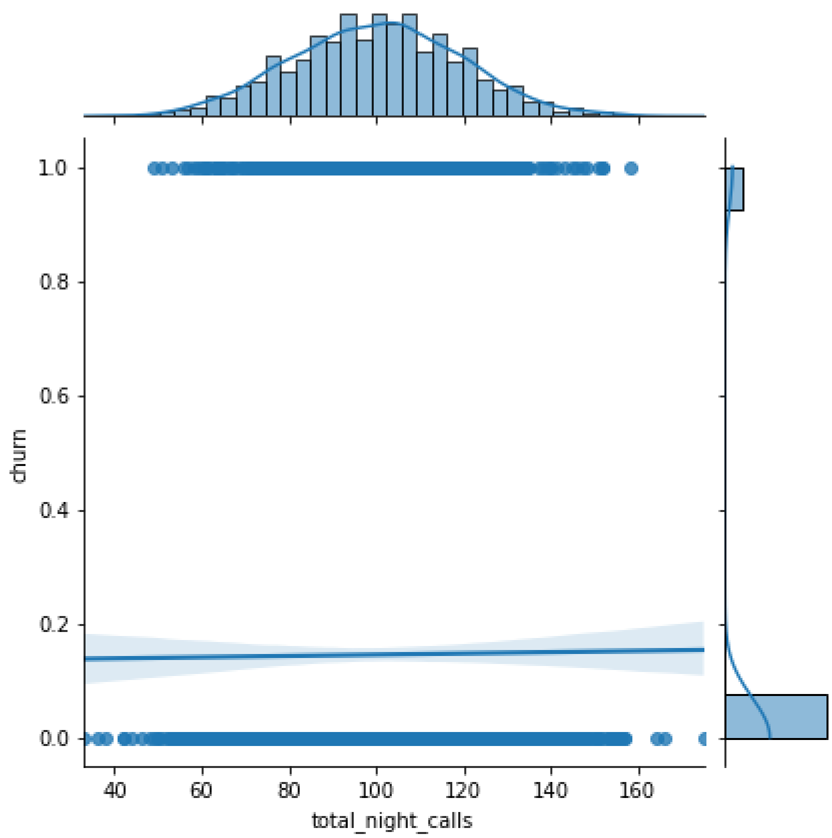

1.0
0.8
0.6
0.4
0.2
0.0
churn
0
10
20
30
40
50
total_day_charge

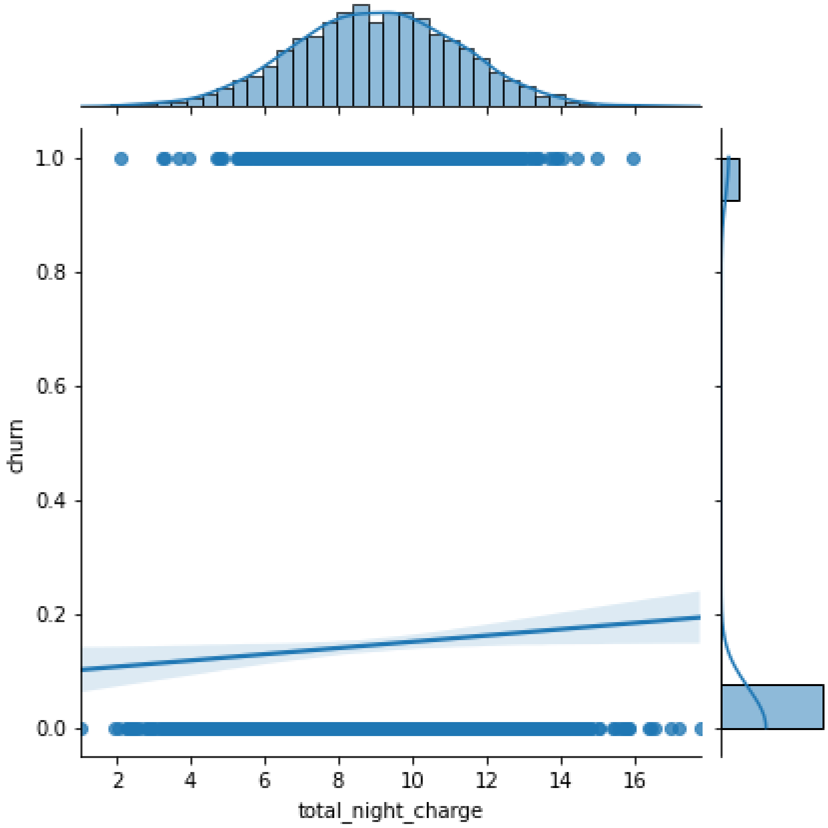

1.0
0.8
0.6
0.4
0.2
0.0
churn
0
50
100
150
200
250
300
350
total_eve_minutes

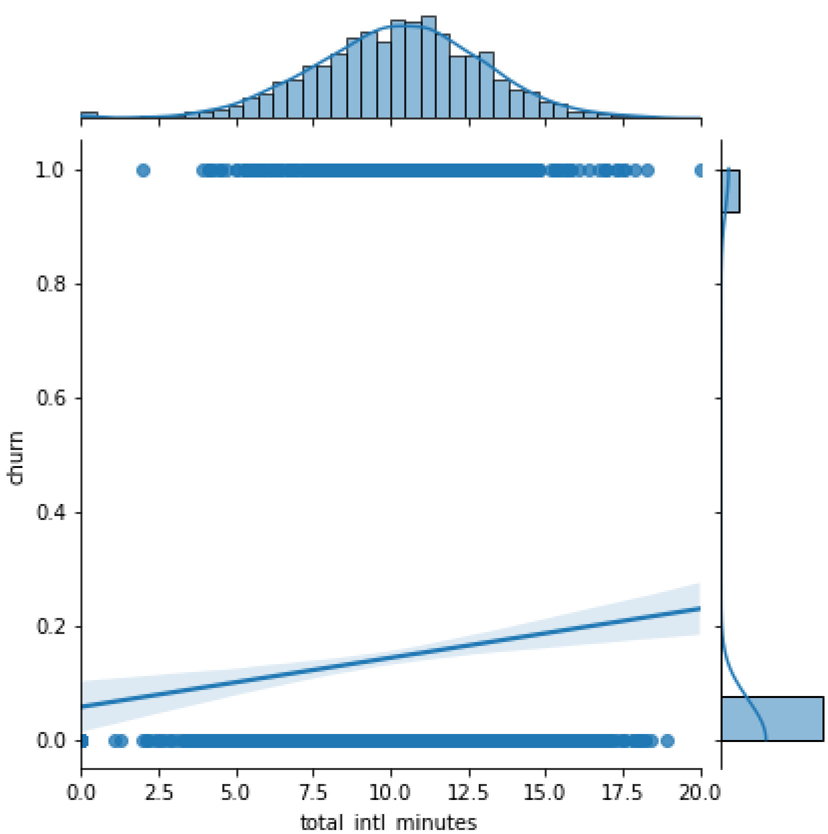

1.0
0.8
0.6
0.4
0.2
0.0
churn
0
20
40
60
80
100
120
140
160
total_eve_calls

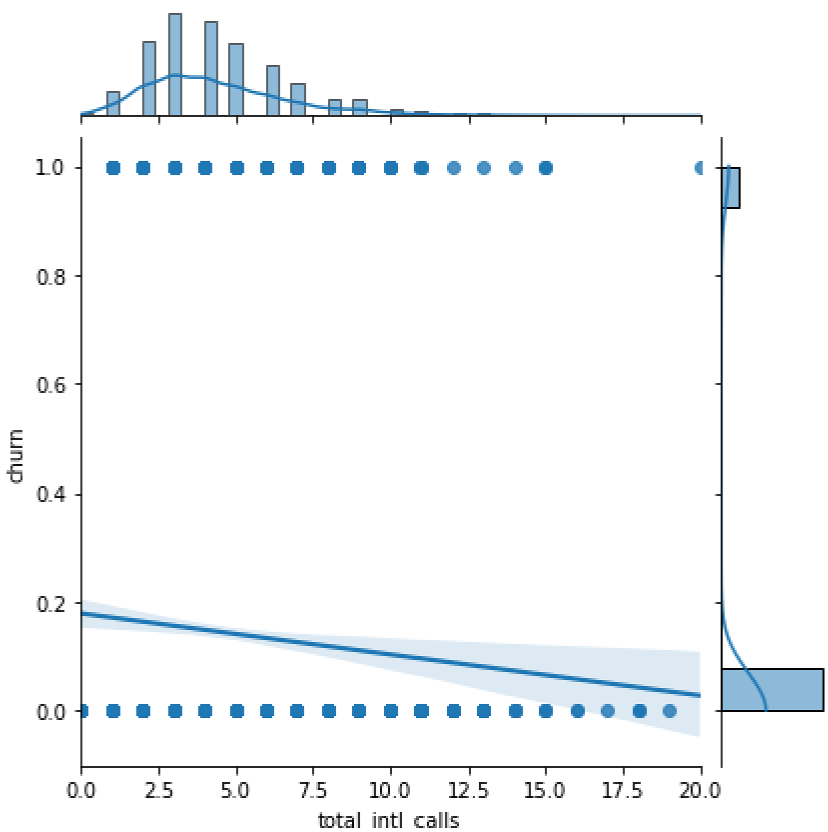

1.0
0.8
0.6
0.4
0.2
0.0
churn
0
5
10
15
20
25
30
total_eve_charge

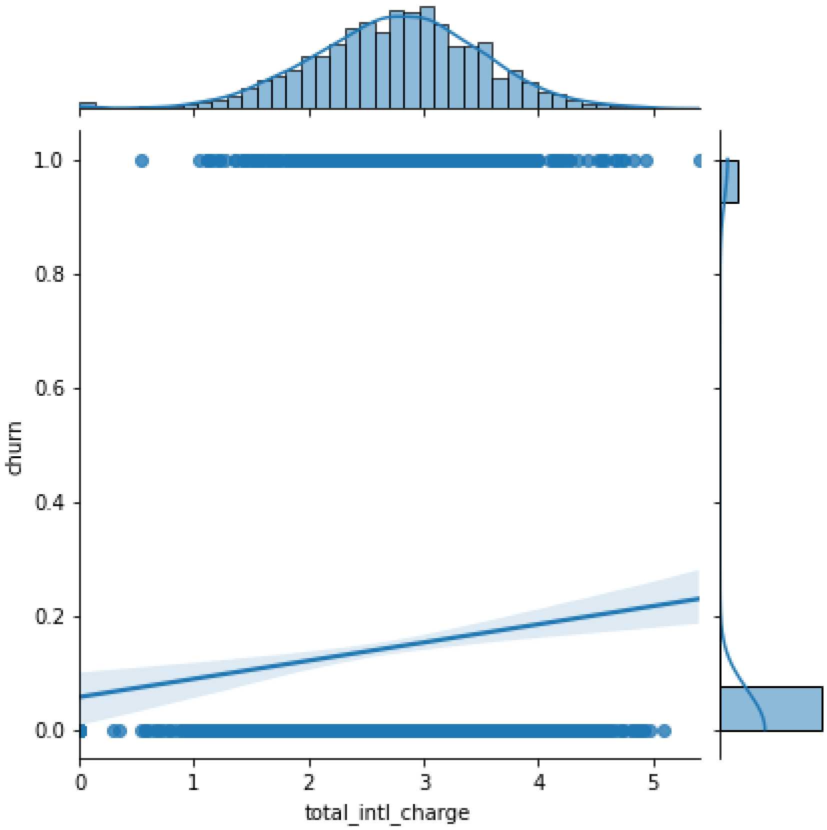

1.0
0.8
0.6
0.4
0.2
0.0
churn
50
100
150
200
250
300
350
total_night_minutes


1.0
0.8
0.6
0.4
0.2
0.0
churn
40
60
80
100
120
140
160
total_night_calls

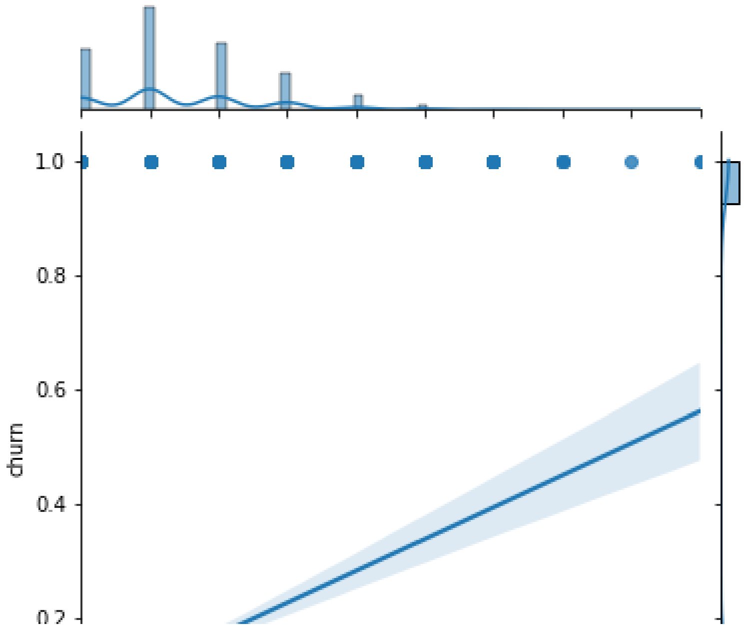

1.0
0.8
0.6
0.4
0.2
0.0
churn
2
4
6
8
10
12
14
16
total_night_charge

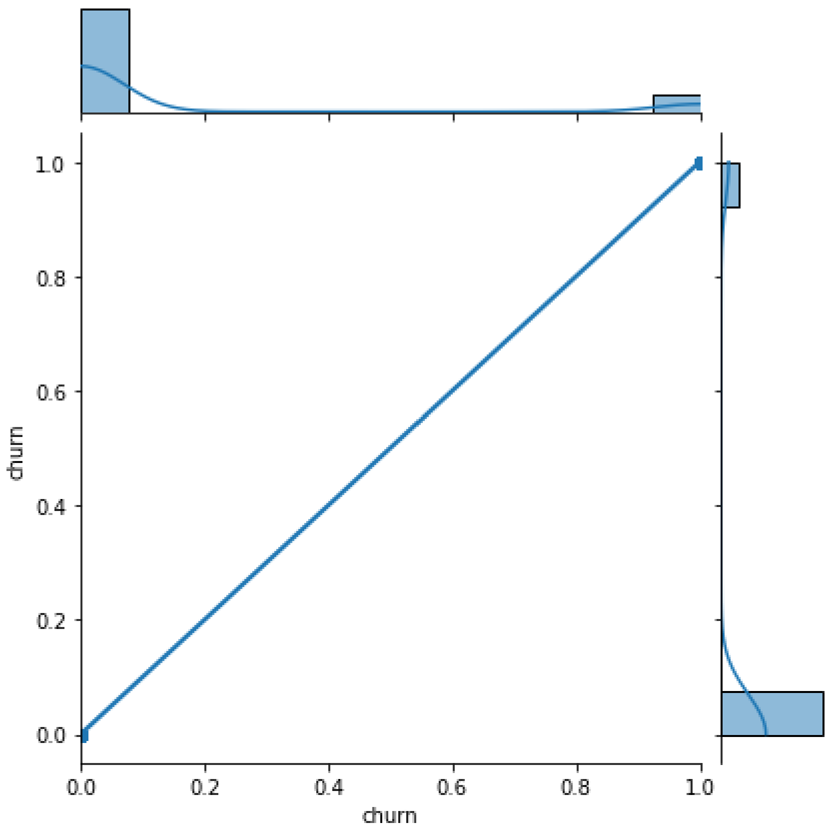

1.0
0.8
0.6
0.4
0.2
0.0
churn
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
total_intl_minutes

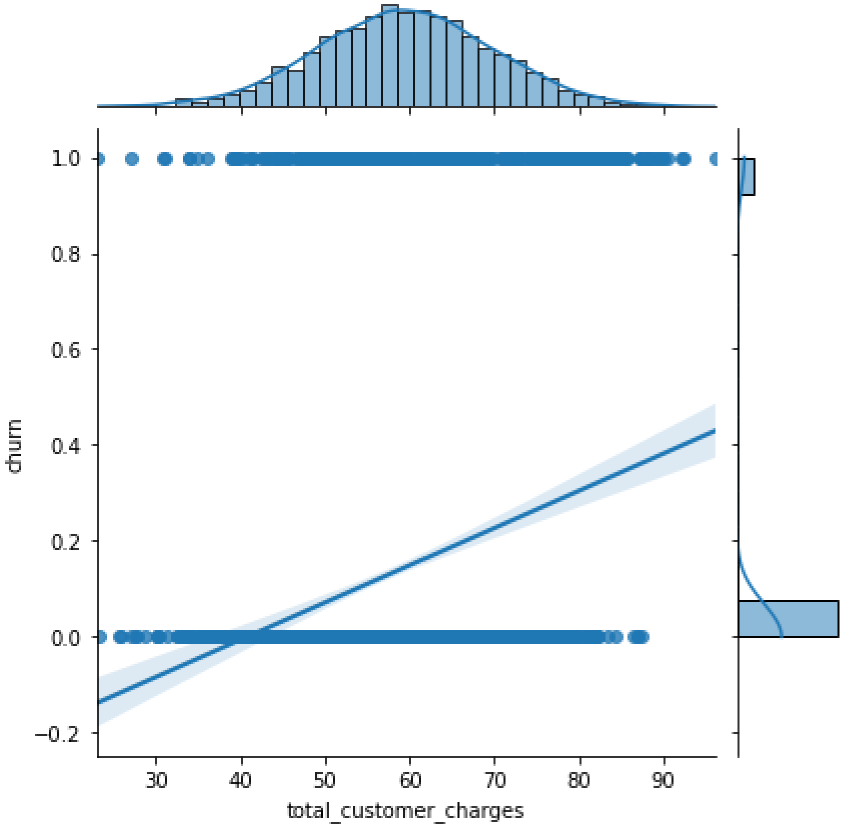

1.0
0.8
0.6
0.4
0.2
0.0
churn
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0
total_intl_calls



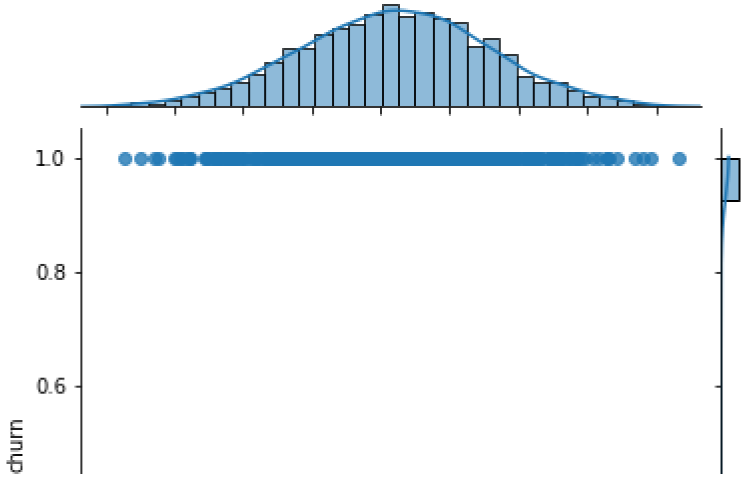
1.0
0.8
0.6
0.4
0.2
0.0
churn
0
1
2
3
4
5
total_intl_charge

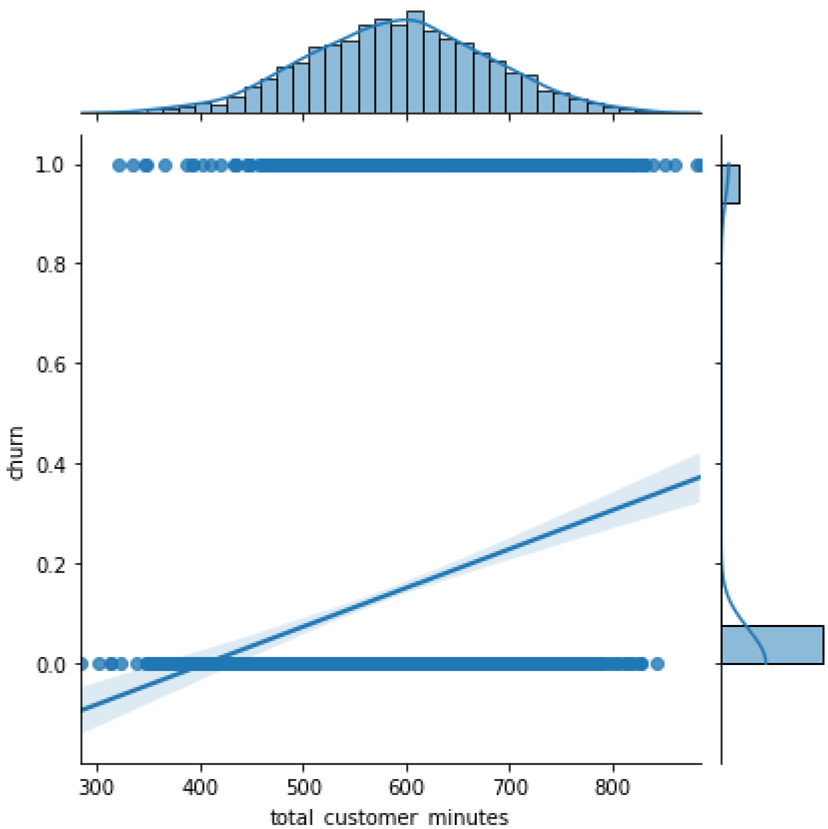

1.0
0.8
0.6
0.4
0.2
churn


1.0
0.8
0.6
0.4
0.2
0.0
churn
0.0
0.2
0.4
0.6
0.8
1.0
churn



1.0
0.8
0.6
0.4
0.2
0.0
−0.2
churn
30
40
50
60
70
80
90
total_customer_charges





Observations

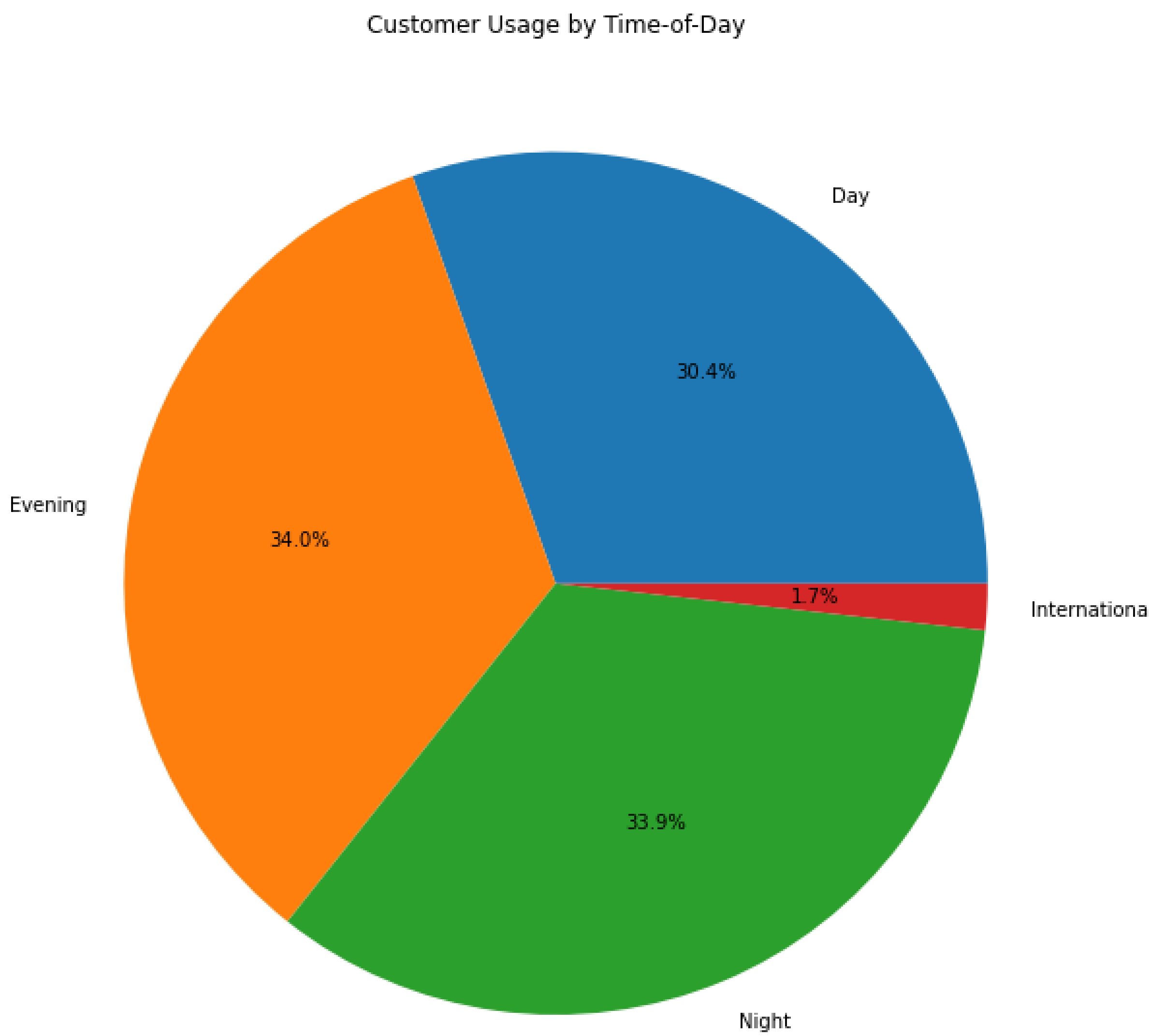
- Customers with international plan seem to slightly lean towards churning
- Customers with voicemail plan seem to slightly lean towards NOT churning
- More voicemail msgs (suggesting high plan usage) slightly leans towards NOT churning
- As a general trend, the **more** calls/minutes/charges that customers use/receives, the **more** they are inclined to churn from SyriaTel.

Let's see how total how much time-of-day affects customer usage

- Day: 5 AM to 5 PM
- Evening: 5 PM to 9 PM
- Night: 9 PM to 4 AM

In [14]:
```
# Set minutes data up for pie chart
data_time = df[['total_day_minutes','total_eve_minutes','total_night_minutes','tc
data_time.rename(index={"total_day_minutes": "Day", "total_eve_minutes": "Evening

# Plot
pie, ax = plt.subplots(figsize=[10,20])
labels = data_time.index
plt.pie(data_time, labels=labels, autopct='%1.1f%%')
plt.title('Customer Usage by Time-of-Day')
plt.show()
```

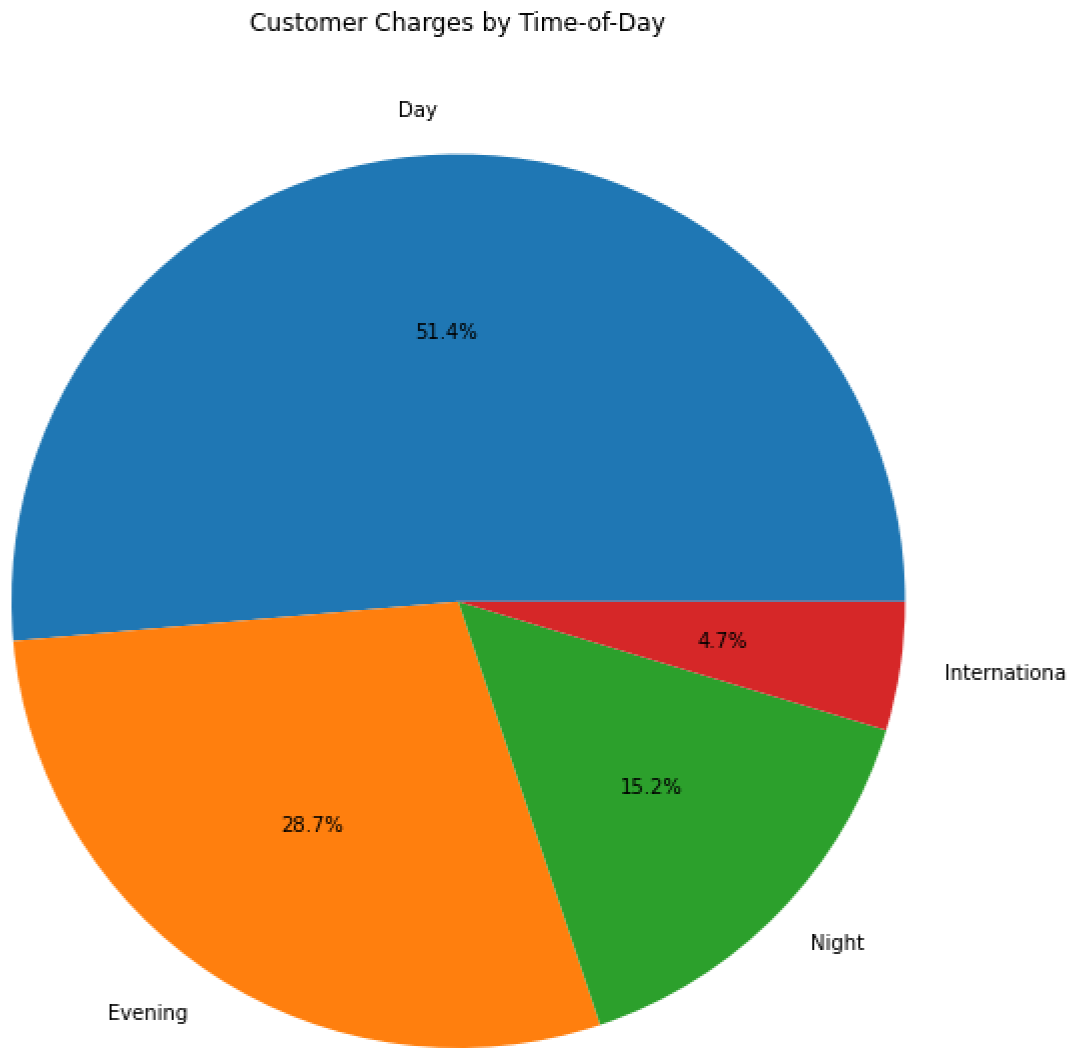
executed in 139ms, finished 01:07:34 2021-04-25



- Surprisingly, each part of the day is equally spent on calls
- Customers spend most of their calls in the latter part of the day (Evening + Night)

In [15]:
```
# Set data up for pie chart
data_charges = df[['total_day_charge','total_eve_charge','total_night_charge','to
data_charges.rename(index={"total_day_charge": "Day", "total_eve_charge": "Evenir

# Plot
pie, ax = plt.subplots(figsize=[10,20])
labels = data_charges.index
plt.pie(data_charges, labels=labels, autopct='%1.1f%%')
plt.title('Customer Charges by Time-of-Day')
plt.show()
```

executed in 123ms, finished 01:07:34 2021-04-25

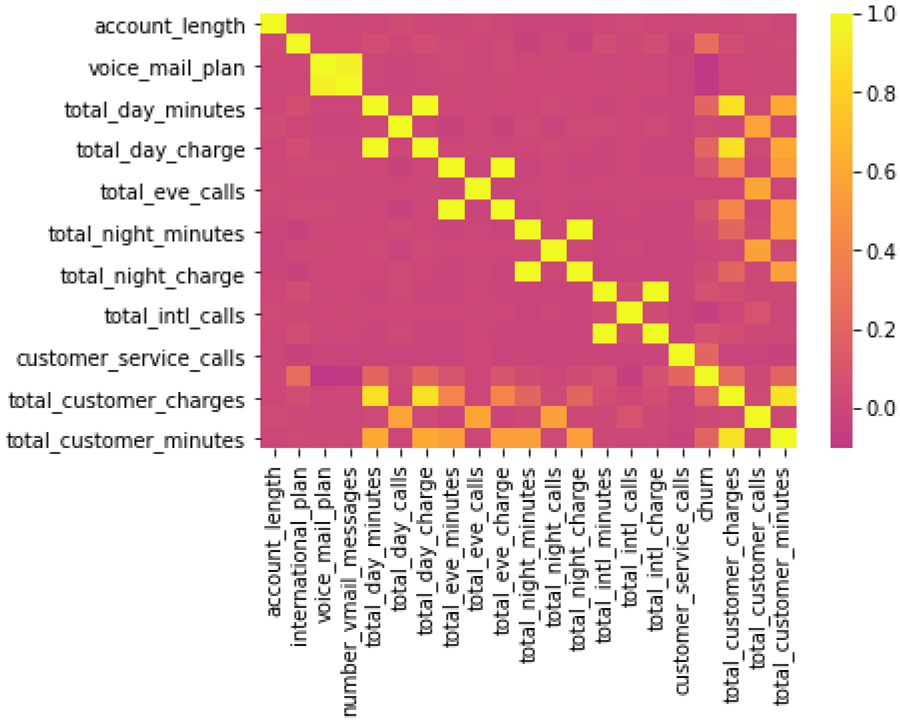


- Despite most usages being at Night and Evening, the total customer charges are the largest for day.

- It may be better to reduce the Day time charges and increase the Evening and Night time charges as it will lead to larger profits.

In [16]:
```
# Heatmap for correlation values
import seaborn as sns
sns.heatmap(df.corr(), cmap='plasma', center=0);
```

executed in 510ms, finished 01:07:35 2021-04-25



There appears to be be some heavy multicollinearity between some factors. Let's identify which ones are causing an issue

In [17]:
```
# Displays if correlation coefficient values is greater than 0.75
df.corr()
abs(df.corr()) > 0.75

# Finds which column pairs have a CC values > 0.75
df_mc = df.corr().abs().stack().reset_index().sort_values(0, ascending=False)

df_mc['pairs'] = list(zip(df_mc.level_0, df_mc.level_1))

df_mc.set_index(['pairs'], inplace = True)

df_mc.drop(columns=['level_1', 'level_0'], inplace = True)

# cc for correlation coefficient
df_mc.columns = ['cc']

df_mc.drop_duplicates(inplace=True)

df_mc[(df_mc.cc>.75) & (df_mc.cc<1)]
```

executed in 46ms, finished 01:07:35 2021-04-25

Out[17]:

pairs	cc
(total_day_charge, total_day_minutes)	1.000000
(total_eve_minutes, total_eve_charge)	1.000000
(total_night_charge, total_night_minutes)	0.999999
(total_intl_minutes, total_intl_charge)	0.999993
(voice_mail_plan, number_vmail_messages)	0.956927
(total_customer_charges, total_customer_minutes)	0.890804
(total_customer_charges, total_day_charge)	0.884757
(total_customer_charges, total_day_minutes)	0.884754

- Let's get rid of the factors related to minutes and keep the charge factors. Price is more important to our overall analysis
- We will also remove number_vmail_messages as the more important factor is that they have a voicemail plan.

In [18]:
```
# Dropping total_day_minutes, total_eve_minutes, total_night_minutes, total_intl_
df.drop(['total_day_minutes','total_eve_minutes','total_night_minutes','total_int
```

executed in 14ms, finished 01:07:35 2021-04-25

In [19]:
```
# Final Cleaned data
df.head()
```
executed in 29ms, finished 01:07:35 2021-04-25

Out[19]:

	state	account_length	international_plan	voice_mail_plan	total_day_calls	total_day_charge	tota
0	KS	128	0	1	110	45.07	
1	OH	107	0	1	123	27.47	
2	NJ	137	0	0	114	41.38	
3	OH	84	1	0	71	50.90	
4	OK	75	1	0	113	28.34	

1.6 Modeling

Now that we have explored the cleaned data, we can finally move on to create models to properly see the effects of each of the factors on telecom customer churning.

1.6.1 Model Training/Test data

In [20]:
```
# Create categories
df_dummy = pd.get_dummies(df)
df_dummy.head()
```
executed in 30ms, finished 01:07:35 2021-04-25

Out[20]:

	account_length	international_plan	voice_mail_plan	total_day_calls	total_day_charge	total_eve_c
0	128	0	1	110	45.07	
1	107	0	1	123	27.47	
2	137	0	0	114	41.38	
3	84	1	0	71	50.90	
4	75	1	0	113	28.34	

5 rows × 66 columns

In [21]:
```
# Create X,y and train/test
X = df_dummy.drop(columns=['churn'], axis=1)
y = df_dummy['churn']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.20, rando
```
executed in 13ms, finished 01:07:35 2021-04-25

1.6.2 Class Imbalance

In [22]:
```
df['churn'].value_counts()
```
executed in 14ms, finished 01:07:35 2021-04-25

Out[22]:
```
0    2850
1     483
Name: churn, dtype: int64
```

In [23]:
```
# SMOTE
smote = SMOTE()
X_train_sm, y_train_sm = smote.fit_resample(X_train, y_train)
```
executed in 43ms, finished 01:07:35 2021-04-25

1.6.3 Logistic Regression

In [24]:
```
# Logistic Regression model
logreg = LogisticRegression(fit_intercept=False, C=1e12, solver='liblinear')
model_log = logreg.fit(X_train_sm, y_train_sm)
```
executed in 92ms, finished 01:07:35 2021-04-25

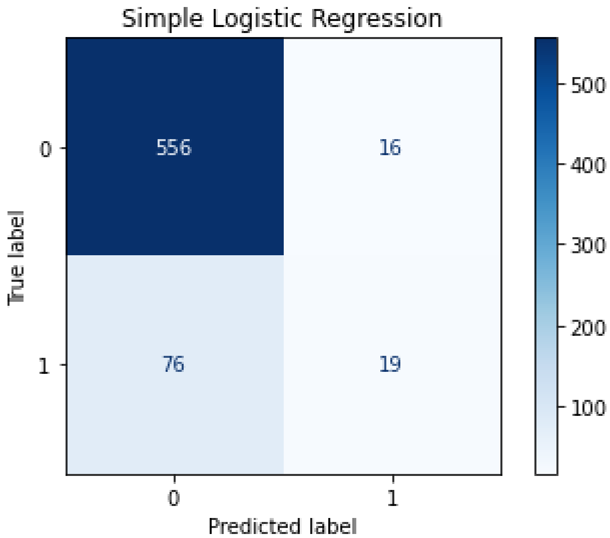
In [25]:
```
# Prediction
y_hat_train = logreg.predict(X_train_sm)
y_hat_test = logreg.predict(X_test)
```
executed in 13ms, finished 01:07:35 2021-04-25

In [26]:
```
display(model_log.score(X_train_sm, y_train_sm))
display(model_log.score(X_test, y_test))
```
executed in 29ms, finished 01:07:35 2021-04-25

```
0.9218612818261633

0.8620689655172413
```

In [27]:
```
# Plot confusion matrix
plot_confusion_matrix(model_log, X_test, y_test,cmap=plt.cm.Blues)
plt.title('Simple Logistic Regression')
plt.show()
```

executed in 187ms, finished 01:07:35 2021-04-25



Type *Markdown* and LaTeX: α^2

In [28]:
```
# we compute our validation metric, recall

print('Training Precision: ', precision_score(y_train_sm, y_hat_train))
print('Testing Precision: ', precision_score(y_test, y_hat_test))
print('\n\n')

print('Training Recall: ', recall_score(y_train_sm, y_hat_train))
print('Testing Recall: ', recall_score(y_test, y_hat_test))
print('\n\n')

print('Training Accuracy: ', accuracy_score(y_train_sm, y_hat_train))
print('Testing Accuracy: ', accuracy_score(y_test, y_hat_test))
print('\n\n')

print('Training F1-Score: ', f1_score(y_train_sm, y_hat_train))
print('Testing F1-Score: ', f1_score(y_test, y_hat_test))
```

executed in 29ms, finished 01:07:35 2021-04-25

```
Training Precision:  0.9674124513618677
Testing Precision:  0.5428571428571428



Training Recall:  0.8731343283582089
Testing Recall:  0.2



Training Accuracy:  0.9218612818261633
Testing Accuracy:  0.8620689655172413



Training F1-Score:  0.9178587909552376
Testing F1-Score:  0.29230769230769227
```

- We want to focus on **Recall** as our main model metric because we want to focus on predicting whether or not a customer is going to churn
- Recall is 87% for Training and 20% for Test. The test score is extremely low. A new model classifier will be needed
- The training score is higher than the recall score, thus is **NOT Overfitting**.

1.6.4 Random Forest

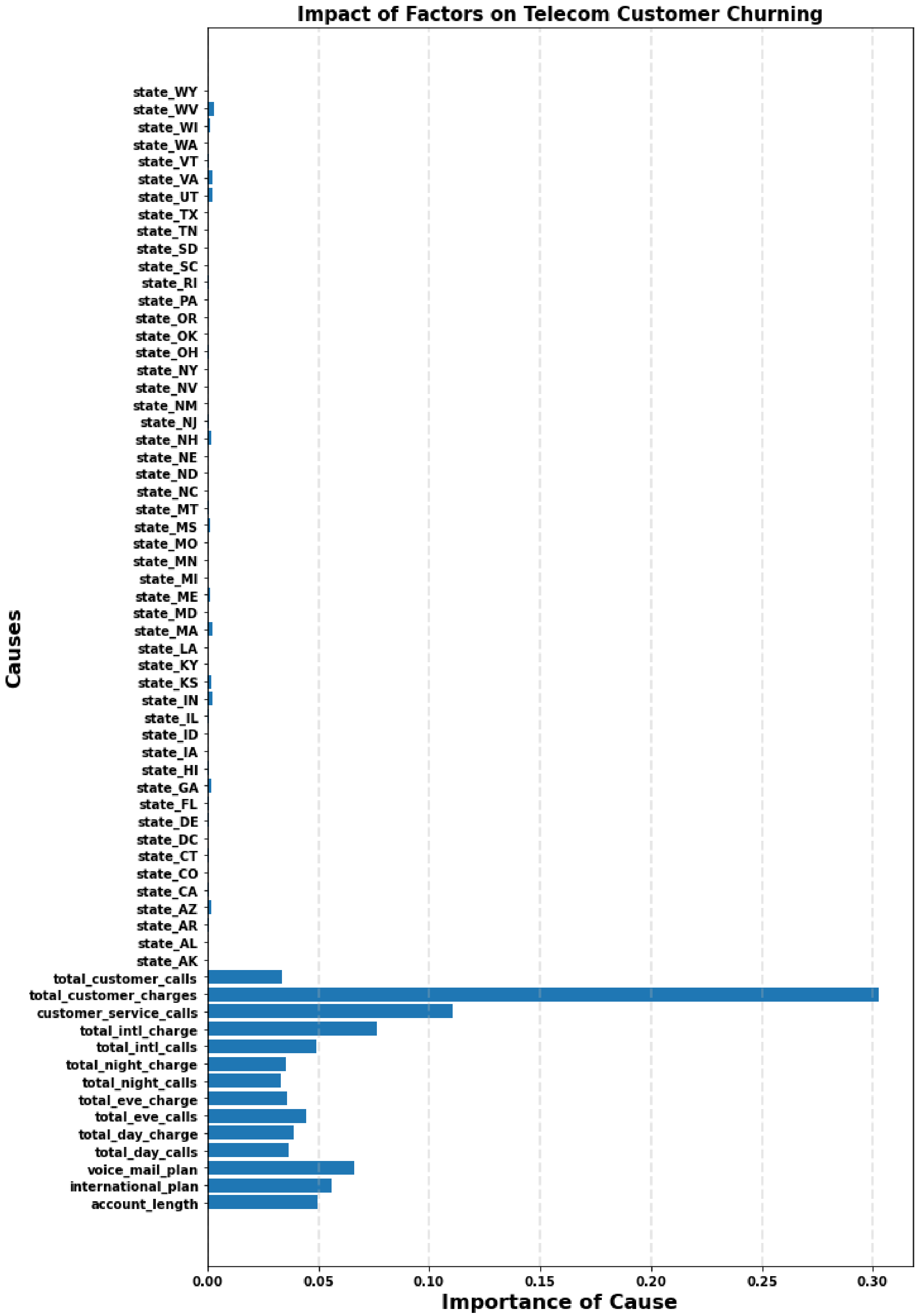
In [29]:
```
# Set up initial forest
tree_clf = DecisionTreeClassifier()
tree_clf.fit(X_train_sm, y_train_sm)
```

executed in 62ms, finished 01:07:35 2021-04-25

Out[29]:
```
DecisionTreeClassifier()
```

In [30]:
```
# Plotting feature importance of models
def plot_feature_importances(model):
    n_features = X_train_sm.shape[1]
    plt.figure(figsize=(10,18))
    plt.barh(range(n_features), model.feature_importances_)
    plt.xticks(fontsize=10, fontweight='bold')
    plt.yticks(np.arange(n_features), X_train_sm.columns.values, fontsize=10, for
    plt.xlabel('Importance of Cause', fontsize=16, fontweight='bold' )
    plt.ylabel('Causes',fontsize=16, fontweight='bold')
    plt.grid(linestyle='--', linewidth=2, axis='x', alpha=0.3)
    plt.title("Impact of Factors on Telecom Customer Churning", fontsize=15, font

plot_feature_importances(tree_clf)
```

executed in 1.68s, finished 01:07:37 2021-04-25


Impact of Factors on Telecom Customer Churning
Causes
Importance of Cause
state_WY
state_WV
state_WI
state_WA
state_VT
state_VA
state_UT
state_TX
state_TN
state_SD
state_SC
state_RI
state_PA
state_OR
state_OK
state_OH
state_NY
state_NV
state_NM
state_NJ
state_NH
state_NE
state_ND
state_NC
state_MT
state_MS
state_MO
state_MN
state_MI
state_ME
state_MD
state_MA
state_LA
state_KY
state_KS
state_IN
state_IL
state_ID
state_IA
state_HI
state_GA
state_FL
state_DE
state_DC
state_CT
state_CO
state_CA
state_AZ
state_AR
state_AL
state_AK
total_customer_calls
total_customer_charges
customer_service_calls
total_intl_charge
total_intl_calls
total_night_charge
total_night_calls
total_eve_charge
total_eve_calls
total_day_charge
total_day_calls
voice_mail_plan
international_plan
account_length
0.00
0.05
0.10
0.15
0.20
0.25
0.30

In [31]:
```
# Test set predictions
val = tree_clf.predict(X_train_sm)
pred = tree_clf.predict(X_test)

# Confusion matrix and classification report
print(confusion_matrix(y_test, pred))
print(classification_report(y_test, pred))
print('Training Recall: ', recall_score(y_train_sm, val)*100)
print('Testing Recall: ', recall_score(y_test, pred)*100)
```
executed in 29ms, finished 01:07:37 2021-04-25

```
[[475  97]
 [ 18  77]]
              precision    recall  f1-score   support

           0       0.96      0.83      0.89       572
           1       0.44      0.81      0.57        95

    accuracy                           0.83       667
   macro avg       0.70      0.82      0.73       667
weighted avg       0.89      0.83      0.85       667

Training Recall:  100.0
Testing Recall:  81.05263157894737
```

- Not Overfitting

In [32]:
```
# Random Forest
forest = RandomForestClassifier()
forest.fit(X_train_sm, np.ravel(y_train_sm))
```
executed in 781ms, finished 01:07:38 2021-04-25

Out[32]:
```
RandomForestClassifier()
```

In [33]:
```
# Finding the best parameters
param_grid = {
    'max_depth': [2, 5, 10, 25],
    'min_samples_split': [2, 5, 10, 20]
}

gs_tree = GridSearchCV(forest, param_grid, cv=3, scoring='recall')
gs_tree.fit(X_train_sm, np.ravel(y_train_sm))

gs_tree.best_params_
```
executed in 18.7s, finished 01:07:57 2021-04-25

Out[33]:
```
{'max_depth': 25, 'min_samples_split': 2}
```

In [34]:
```
# Applying best Random Forest parameters
forest = RandomForestClassifier(max_depth=25, min_samples_split=2)
forest.fit(X_train_sm, np.ravel(y_train_sm))

vals = forest.predict(X_train_sm)
preds = forest.predict(X_test)

print(confusion_matrix(y_test, preds))
print(classification_report(y_test, preds))
print('Training Recall: ', recall_score(y_train_sm, vals)*100)
print('Testing Recall: ', recall_score(y_test, preds)*100)
```

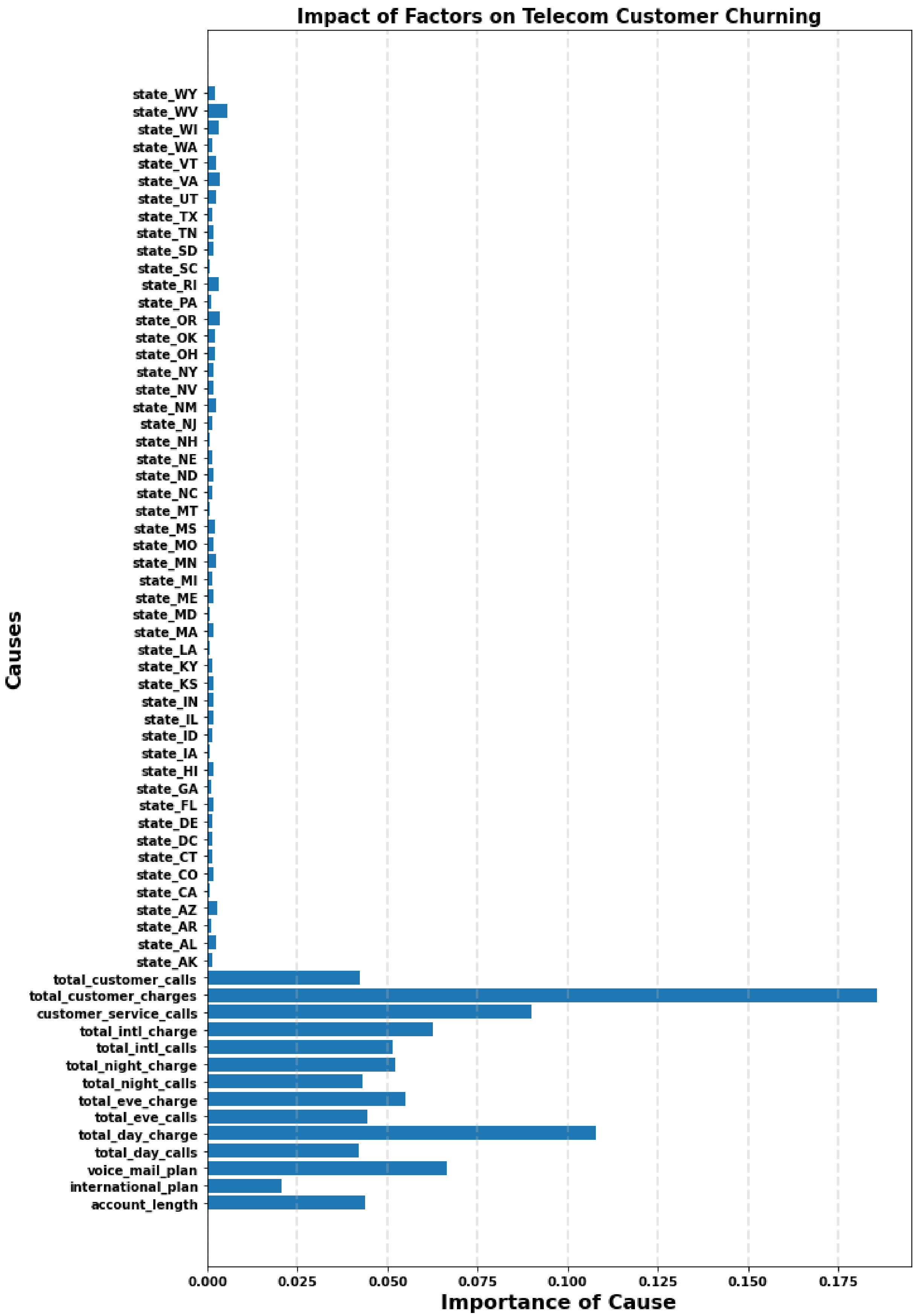
executed in 872ms, finished 01:07:58 2021-04-25

```
[[552  20]
 [ 21  74]]
              precision    recall  f1-score   support

           0       0.96      0.97      0.96       572
           1       0.79      0.78      0.78        95

    accuracy                           0.94       667
   macro avg       0.88      0.87      0.87       667
weighted avg       0.94      0.94      0.94       667

Training Recall:  100.0
Testing Recall:  77.89473684210526
```

- Our Confusion matrix is looking better than before. Predicting more True churns than False churns
- Training Recall score is higher than test. **Not Overfitting**
- Testing Recall score is still low (76%).
- Should keep looking for a better Classifier

In [35]:
```
plot_feature_importances(forest)
```

executed in 1.71s, finished 01:07:59 2021-04-25



- The individual states do not seem to have a large influence on the customer churn
- **Total customer charges_** and **total day charge** appears to have the largest influence on churn

1.6.5 XG Boost

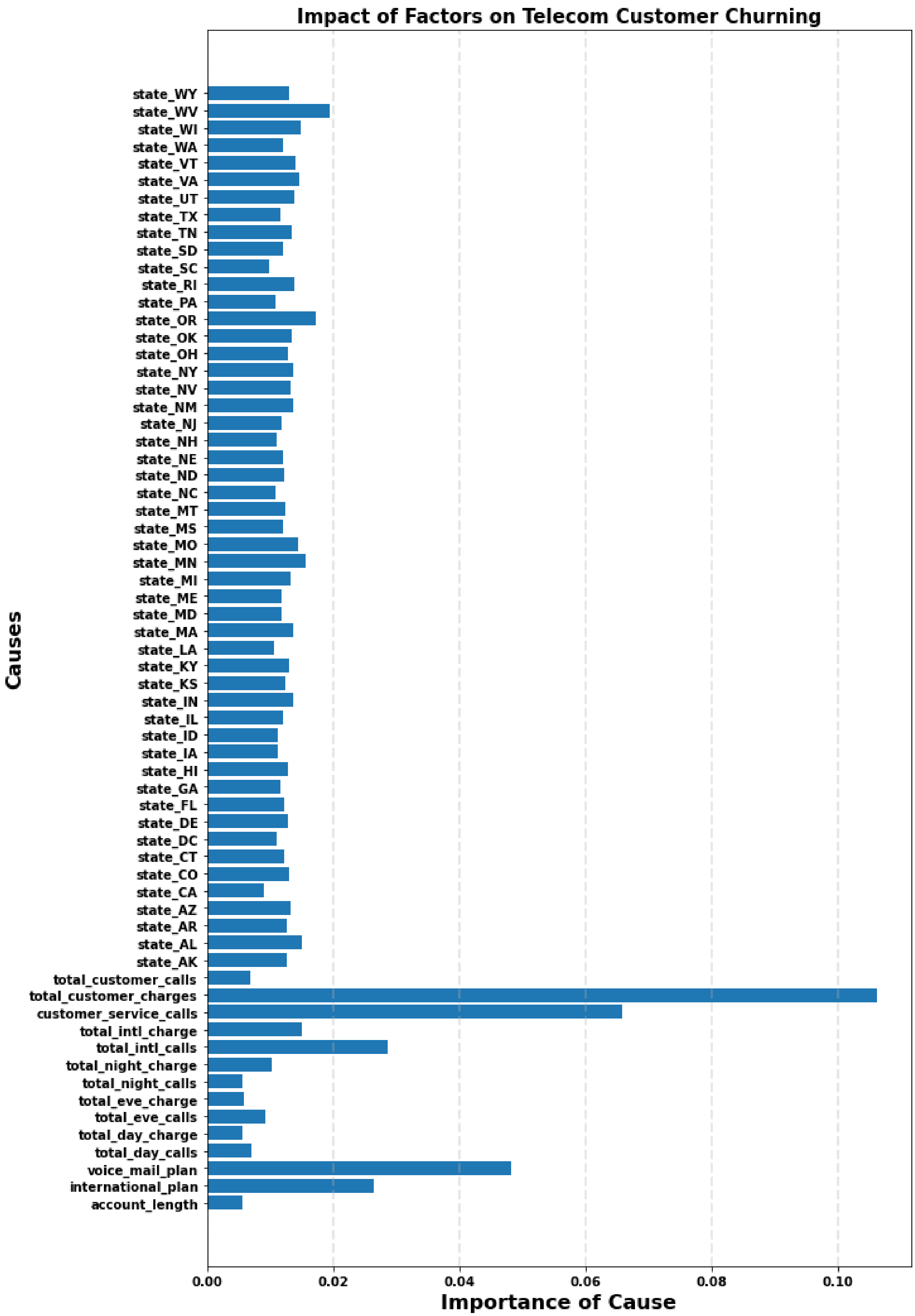
In [36]:
```
# Instantiate XGBClassifier
xgb = XGBClassifier()

# Fit XGBClassifier
xgb.fit(X_train_sm, np.ravel(y_train_sm))

# Predict on training and test sets
training_preds = xgb.predict(X_train_sm)
test_preds = xgb.predict(X_test)

# Accuracy of training and test sets
training_recall = recall_score(y_train_sm, training_preds)
test_recall = recall_score(y_test, test_preds)

print('Training Recall: {:.4}%'.format(training_recall * 100))
print('Test Recall: {:.4}%'.format(test_recall * 100))
```

executed in 622ms, finished 01:08:00 2021-04-25

```
Training Recall: 98.95%
Test Recall: 82.11%
```

In [37]:
```
plot_feature_importances(xgb)
```

executed in 1.71s, finished 01:08:02 2021-04-25



In [38]:
```
param_grid = {
    'learning_rate': [0.1, 0.2],
    'max_depth': [6],
    'min_child_weight': [1, 2],
    'subsample': [0.5, 0.7],
    'n_estimators': [100],
}

grid_clf = GridSearchCV(xgb, param_grid, scoring='recall')
grid_clf.fit(X_train_sm, np.ravel(y_train_sm))

best_parameters = grid_clf.best_params_

print('Grid Search found the following optimal parameters: ')
for param_name in sorted(best_parameters.keys()):
    print('%s: %r' % (param_name, best_parameters[param_name]))

training_preds = grid_clf.predict(X_train_sm)
test_preds = grid_clf.predict(X_test)

training_recall = recall_score(y_train_sm, training_preds)
test_recall = recall_score(y_test, test_preds)

print('')
print('Training Recall: {:.4}%'.format(training_recall * 100))
print('Test Recall: {:.4}%'.format(test_recall * 100))
```

executed in 13.1s, finished 01:08:15 2021-04-25

```
Grid Search found the following optimal parameters: 
learning_rate: 0.2
max_depth: 6
min_child_weight: 1
n_estimators: 100
subsample: 0.7

Training Recall: 97.67%
Test Recall: 81.05%
```

In [39]:
```
print(confusion_matrix(y_test, test_preds))
print(classification_report(y_test, test_preds))
print('')
print('Training Recall: {:.4}%'.format(training_recall * 100))
print('Test Recall: {:.4}%'.format(test_recall * 100))
```

executed in 28ms, finished 01:08:15 2021-04-25

```
[[561  11]
 [ 18  77]]
              precision    recall  f1-score   support

           0       0.97      0.98      0.97       572
           1       0.88      0.81      0.84        95

    accuracy                           0.96       667
   macro avg       0.92      0.90      0.91       667
weighted avg       0.96      0.96      0.96       667


Training Recall: 97.67%
Test Recall: 81.05%
```

- Confusion Matrix shows a great True Churn prediction rate compared to False Churn.
- Higher Training recall than Test recall. **Not Overfitting**
- **XG Boost has the highest testing Recall from the previous models.**
- We shall use this for our predictive analysis

1.6.6 Final Model Selection

In [40]:
```
df_final = pd.DataFrame(xgb.feature_importances_, X.columns, columns = ['coeffici
df_final.head()
```

executed in 60ms, finished 01:08:15 2021-04-25

Out[40]:

	coefficient
account_length	0.005636
international_plan	0.026481
voice_mail_plan	0.048244
total_day_calls	0.006943
total_day_charge	0.005589

In [41]:
```
df_final_detail = df_final[:14]
display(df_final_detail.head())
df_final_state = df_final[14:]
display(df_final_state.head())
```
executed in 13ms, finished 01:08:15 2021-04-25

	coefficient
account_length	0.005636
international_plan	0.026481
voice_mail_plan	0.048244
total_day_calls	0.006943
total_day_charge	0.005589

	coefficient
state_AK	0.012553
state_AL	0.015029
state_AR	0.012521
state_AZ	0.013282
state_CA	0.009105

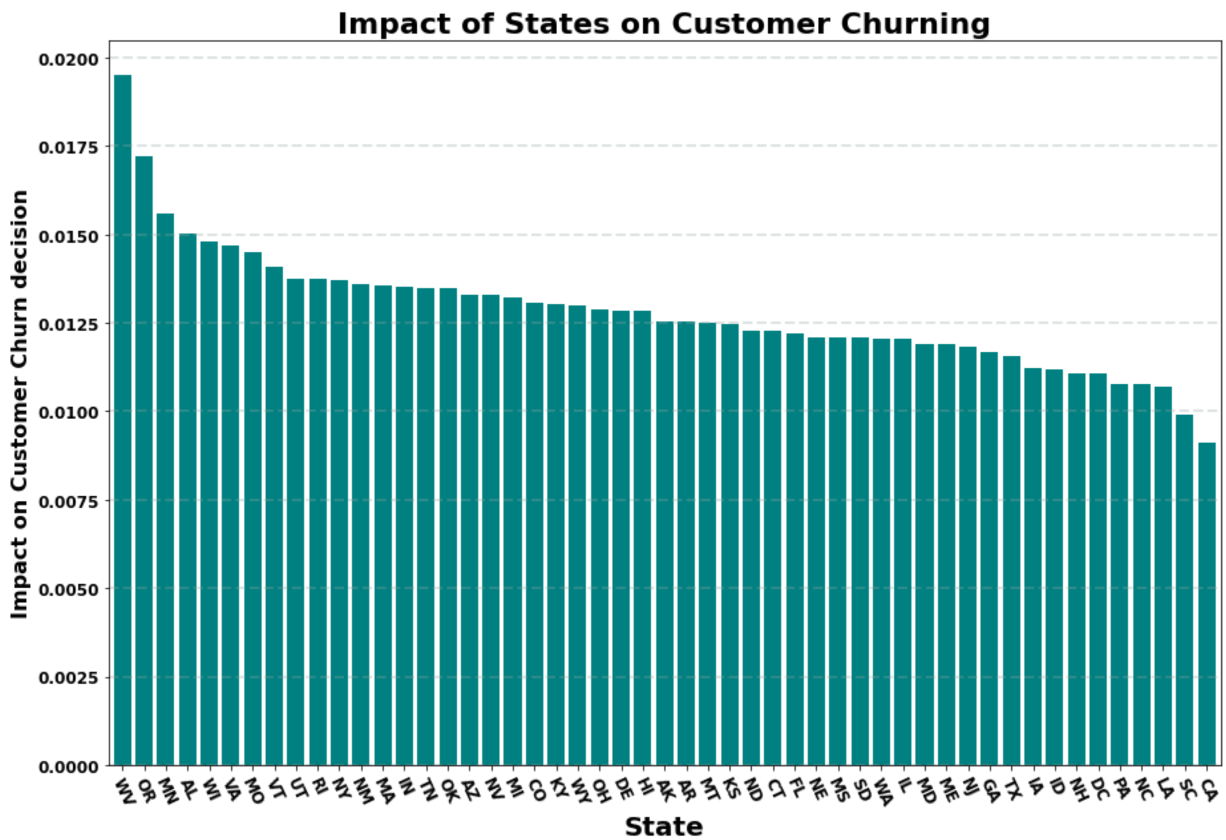
In [42]:
```
# Renaming states indexes
as_list = df_final_state.index.values.tolist()
as_list = [i.replace('state_','') for i in as_list]
df_final_state.index = as_list
df_final_state.head()
```
executed in 14ms, finished 01:08:15 2021-04-25

Out[42]:

	coefficient
AK	0.012553
AL	0.015029
AR	0.012521
AZ	0.013282
CA	0.009105

In [43]:
```
# Create a barplot of state's impact on churn
ax = df_final_state.sort_values(by=['coefficient'], ascending=False).plot(kind='b
ax.set_title("Impact of States on Customer Churning", fontsize = 22, fontweight =
ax.set_xlabel("State", fontsize=20, fontweight='bold')
ax.set_ylabel("Impact on Customer Churn decision", fontsize=16, fontweight='bold'
ax.get_legend().remove()

plt.xticks(rotation=-65, fontsize=12, fontweight='bold')
plt.yticks(fontsize=12, fontweight='bold')
plt.grid(color='#95a5a6', linestyle='--', linewidth=2, axis='y', alpha=0.3)
```

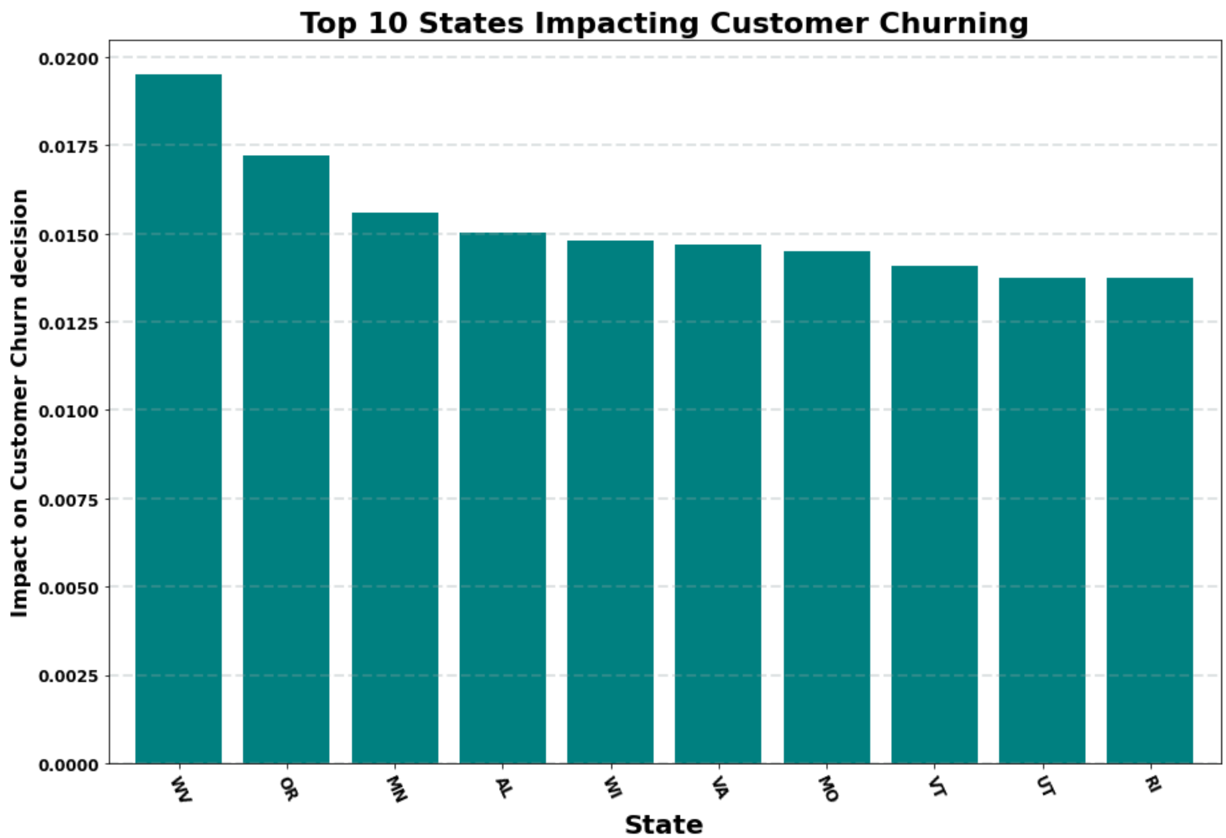
executed in 1.27s, finished 01:08:16 2021-04-25



- There seems to be a slightly larger churn rate in **state_WV**. This may need to be investigated further.
- No apparent correlation can be found with states with higher impact

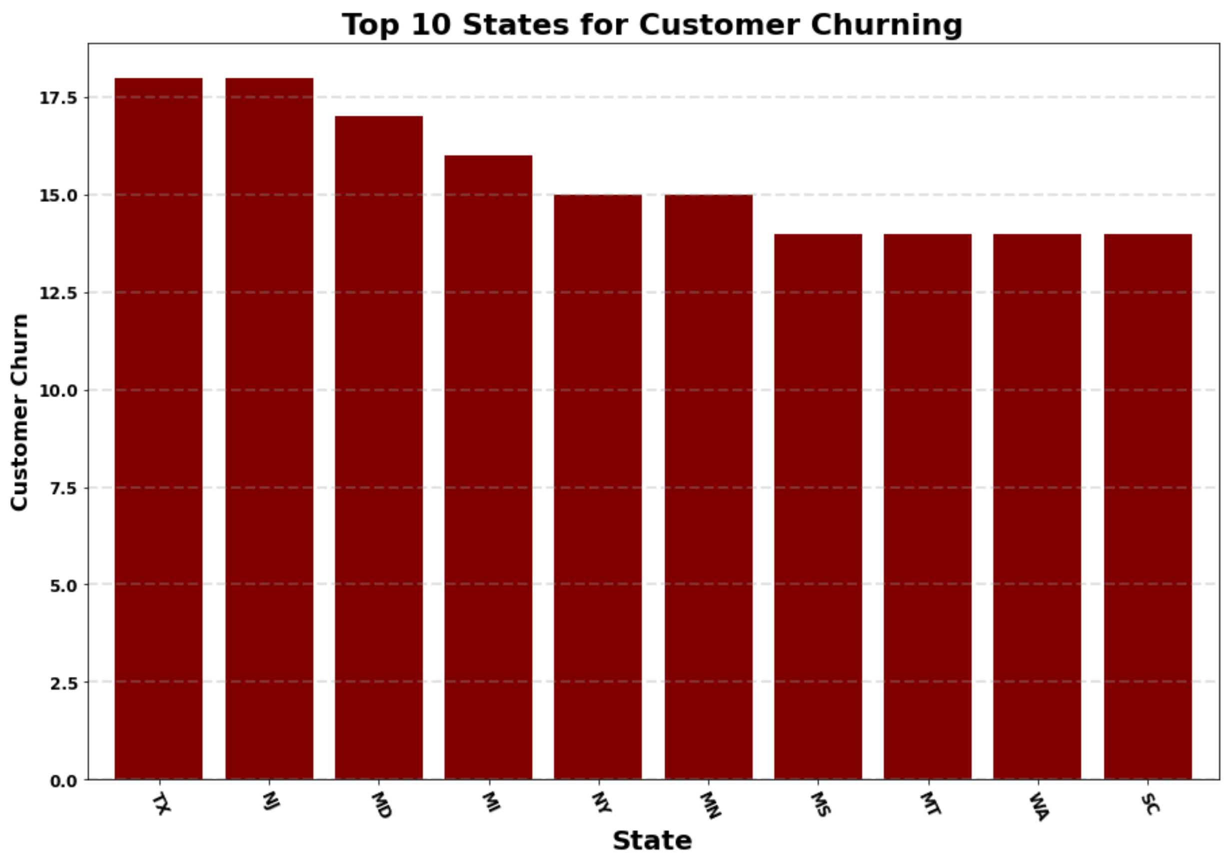
In [44]:
```
# Create a barplot of top 5 state's impact on churn
ax = df_final_state.sort_values(by=['coefficient'], ascending=False).head(10).plo
ax.set_title("Top 10 States Impacting Customer Churning", fontsize = 22, fontweig
ax.set_xlabel("State", fontsize=20, fontweight='bold')
ax.set_ylabel("Impact on Customer Churn decision", fontsize=16, fontweight='bold'
ax.get_legend().remove()

plt.xticks(rotation=-65, fontsize=12, fontweight='bold')
plt.yticks(fontsize=12, fontweight='bold')
plt.grid(color='#95a5a6', linestyle='--', linewidth=2, axis='y', alpha=0.3)
```

executed in 249ms, finished 01:08:16 2021-04-25



In [45]:
```
# Create a barplot of top 5 state's churn based on original dataset
ax = df.groupby(['state']).sum().churn.sort_values(ascending=False).head(10).plot
ax.set_title("Top 10 States for Customer Churning", fontsize = 22, fontweight = '
ax.set_xlabel("State", fontsize=20, fontweight='bold')
ax.set_ylabel("Customer Churn", fontsize=16, fontweight='bold');

plt.xticks(rotation=-65, fontsize=12, fontweight='bold')
plt.yticks(fontsize=12, fontweight='bold')
plt.grid(color='#95a5a6', linestyle='--', linewidth=2, axis='y', alpha=0.3)
```

executed in 219ms, finished 01:08:17 2021-04-25



- Original Data set results for top churn and model feature importance appear to be showing different results
- Will need further investigation on individual state's impact on churn

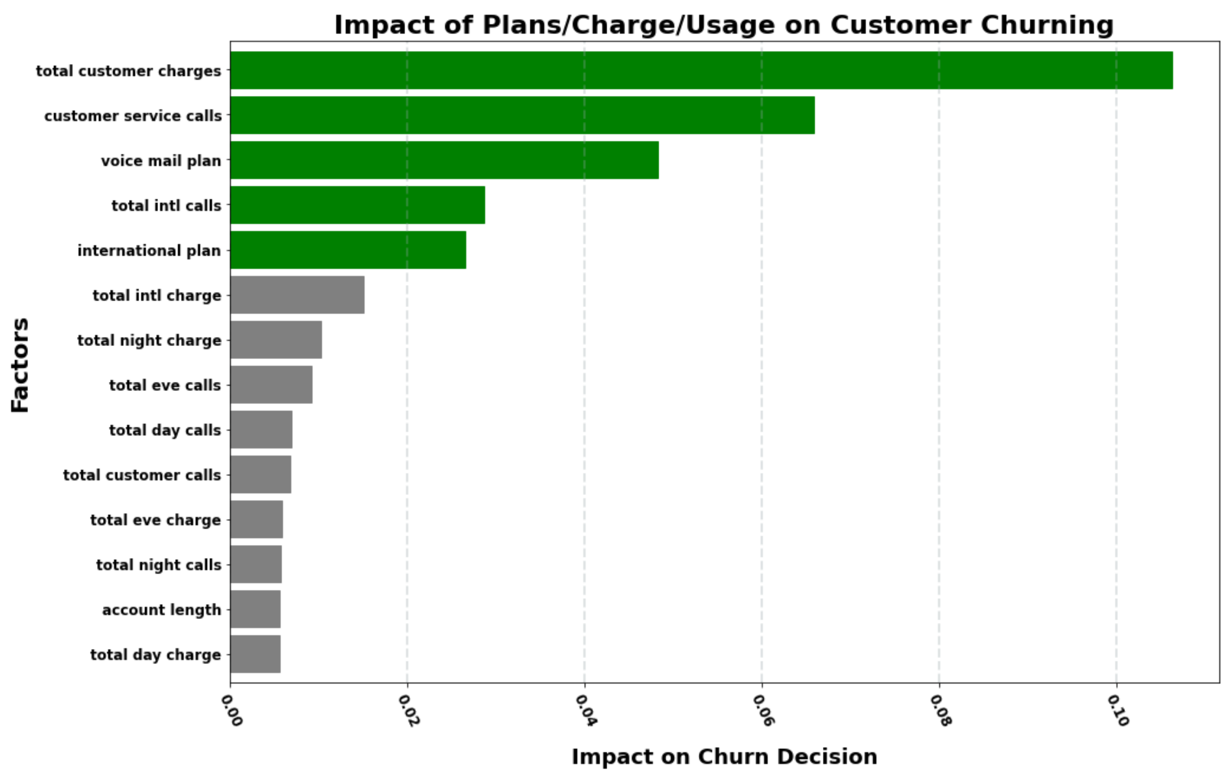
In [46]:
```
# Renaming detail indexes
detail_list = df_final_detail.index.values.tolist()
detail_list = [i.replace('_',' ') for i in detail_list]
df_final_detail.index = detail_list
df_final_detail.head()
```
executed in 14ms, finished 01:08:17 2021-04-25

Out[46]:

	coefficient
account length	0.005636
international plan	0.026481
voice mail plan	0.048244
total day calls	0.006943
total day charge	0.005589

In [47]:
```
# Create a barplot of customer plans/usage's impact on churn
ax = df_final_detail.sort_values(by=['coefficient'], ascending=True).plot(kind='b
ax.set_title("Impact of Plans/Charge/Usage on Customer Churning", fontsize = 22,
ax.set_ylabel("Factors", fontsize=20, fontweight='bold')
ax.set_xlabel("Impact on Churn Decision", fontsize=18, fontweight='bold', labelpa
ax.get_legend().remove()

for column in ax.patches:
    if column.get_width() > 0.02:
        column.set_color('green')
    else:
        column.set_color('grey')

plt.xticks(rotation=-65, fontsize=12, fontweight='bold')
plt.yticks(fontsize=12, fontweight='bold')
plt.grid(color='#95a5a6', linestyle='--', linewidth=2, axis='x', alpha=0.3)
```

executed in 531ms, finished 01:08:17 2021-04-25



- **total_customer_charges** appear to be the largest cause of churn. This makes sense seeing that most customers
- **customer_service_calls** come 2nd influence to the plans. There may need to be an improvement to the customer service section of the company.
- **voice_mail_plan** and **international_plan** are the relatively high factors to customer churn. There may be some issues with how both plans are being handled.

1.7 Conclusions

The analysis of the SyriaTel customer churn dataset resulted in the following conclusions:

- There needs to be an improvement to the **Voice mail plan** and **International plan**. The customer churn is heavily affected by the effectiveness of the plans. These plans need to be further examined to entice customers to stay with SyriaTel.
- The **Customer service** department may need examining. We need to check staffing to see what is causing the customers to turn away from SyriaTel as they submit for help through the customer service line.
- We can predict future customer churn with our final model. This should help us mitigate customer losses if we contact the customer earlier for their input.

1.8 Recommendations

What can we do right now?

- **We can adjust the minute-to-charge rates for day-evening-night.** The high total charges appear to be causing the most impact on SyriaTel customer's churning. Despite most of our customers utilizing around 2/3 of their total minutes in the evening and night, the day-time charges overtake the two charges combined. We may be able to get more customer satisfaction from them knowing that their charges in the day-time are being lowered, despite most of their minutes are spent in the evening. This will allow us to maintain profits from call charges while maintaining a lower churn.
- **Increase the staffing of the Customer Service department.** Increased staffing will readily make available more customers to be directly in line with staff. This will increase customer satisfaction. Additionally, more staffing may help us identify the common issues customers are having more quickly, allowing SyriaTel to address specific issues quickly as well. Overall, customer churn should go down.
- **Reduce the charge of international calls/plan** This will reduce customer churn over the factors of the international plan and its charges. It is not worth losing a customer over a specific plan they have chosen. We should keep the customer to profit off of the general uses outside of the international plan.

1.9 Next Steps

Further analysis of the SyriaTel data could yield additional insights to other recommendations

- **Locate what factors are causing a larger churn impact within specific states**
- **Create an alert system that detects when individual customer are in range of possibly churning**
- **Investigate a change over system from international to domestic plans on customer churn**