

In [1]:

```
#necessary imports
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings('ignore')
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB, BaseEstimator
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import GradientBoostingClassifier, RandomForestClassifier, BaseEnsemble
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix, make_scorer, recall_score
#from sklearn.datasets import make_classification

#from imblearn.over_sampling import SMOTE

import matplotlib.pyplot as plt
import seaborn as sns
```

The most suitable metric to employ in this dataset is Recall. Utilizing Recall is advantageous because it enables the implementation of more effective customer retention strategies. In this context, it is more beneficial to correctly identify a customer as 'exited' and apply retention strategies to keep them engaged, rather than failing to identify a customer who has exited and consequently missing the opportunity to employ retention tactics to ensure their continued subscription to the service.

In [2]:

```
df_train = pd.read_csv('/Users/jamesmaikara/Downloads/training_set.csv', index_col=0)
df_train.head()
```

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	...	total eve calls	total eve charge	total night minutes	total night calls
2682	DC	55	510	354-5058	yes	no	0	106.1	77	18.04	...	100	10.50	96.4	92
3304	IL	71	510	330-7137	yes	no	0	186.1	114	31.64	...	140	16.88	206.5	80
757	UT	112	415	358-5953	no	no	0	115.8	108	19.69	...	111	20.68	184.6	78
2402	NY	77	415	388-9285	no	yes	33	143.0	101	24.31	...	102	18.04	104.9	120
792	NV	69	510	397-6789	yes	yes	33	271.5	98	46.16	...	102	21.54	165.4	85

5 rows x 21 columns



In [3]:

```
df_train.info()

<class 'pandas.core.frame.DataFrame'>
Index: 2999 entries, 2682 to 1061
Data columns (total 21 columns):
#   Column              Non-Null Count  Dtype
---  ---
0   state                2999 non-null   object
```

```

1   account length      2999 non-null    int64
2   area code           2999 non-null    int64
3   phone number        2999 non-null    object
4   international plan   2999 non-null    object
5   voice mail plan      2999 non-null    object
6   number vmail messages 2999 non-null    int64
7   total day minutes    2999 non-null    float64
8   total day calls      2999 non-null    int64
9   total day charge     2999 non-null    float64
10  total eve minutes    2999 non-null    float64
11  total eve calls      2999 non-null    int64
12  total eve charge     2999 non-null    float64
13  total night minutes  2999 non-null    float64
14  total night calls    2999 non-null    int64
15  total night charge   2999 non-null    float64
16  total intl minutes   2999 non-null    float64
17  total intl calls     2999 non-null    int64
18  total intl charge    2999 non-null    float64
19  customer service calls 2999 non-null    int64
20  churn                2999 non-null    bool

```

```
dtypes: bool(1), float64(8), int64(8), object(4)
```

```
memory usage: 495.0+ KB
```

In [4]:

```
df_train['churn'].value_counts()
```

Out[4]:

```

churn
False    2569
True      430
Name: churn, dtype: int64

```

Initial Model

In [5]:

```

#functions are designed to enhance the data transformation process and improve the visualization of model performance
def transform_df(df):
    df['international plan'] = df['international plan'].apply(lambda x: 1 if x.lower() == 'yes' else 0)
    df['voice mail plan'] = df['voice mail plan'].apply(lambda x: 1 if x.lower() == 'yes' else 0)

    return df

def plot_conf_matrix(y_true, y_pred):
    cm = confusion_matrix(y_true, y_pred)
    plt.figure(figsize=(10, 7))
    sns.heatmap(cm, annot=True, cmap=sns.color_palette('Blues_d'), fmt='0.5g', annot_kws={"size": 16})
    plt.xlabel('Predictions')
    plt.ylabel('Actuals')
    plt.ylim([0,2])
    plt.show()

```

In [6]:

```

features_to_use = ['account length', 'international plan', 'voice mail plan', 'number vmail messages',
                  'total day charge', 'total eve charge', 'total night charge', 'total intl charge',
                  'customer service calls']
target = ['churn']

```

In [7]:

```
df_train_transformed = transform_df(df_train)
df_train_transformed.head()
```

Out[7]:

	state	account length	area code	phone number	international plan	voice mail plan	number vmail messages	total day minutes	total day calls	total day charge	...	total eve calls	total eve charge	total night minutes	total night calls
2682	DC	55	510	354-5058	1	0	0	106.1	77	18.04	...	100	10.50	96.4	92
3304	IL	71	510	330-7137	1	0	0	186.1	114	31.64	...	140	16.88	206.5	80
757	UT	112	415	358-5953	0	0	0	115.8	108	19.69	...	111	20.68	184.6	78
2402	NY	77	415	388-9285	0	1	33	143.0	101	24.31	...	102	18.04	104.9	120
792	NV	69	510	397-6789	1	1	33	271.5	98	46.16	...	102	21.54	165.4	85

5 rows x 21 columns



In [8]:

```
#splitting dataset into training and testing sets
X = df_train_transformed[features_to_use]
y = df_train_transformed[target]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=.25, random_state=1)
X_train.shape, X_test.shape
```

Out[8]:

((2249, 9), (750, 9))

In [9]:

```
#using the fit_resample method to perform over-sampling to address class imbalance
from imblearn.over_sampling import SMOTE
smote = SMOTE()
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)
```

In [10]:

```
rf1 = RandomForestClassifier()
rf1.fit(X_train_resampled, y_train_resampled)
```

Out[10]:

▼ RandomForestClassifier

RandomForestClassifier()

In [11]:

```
y_preds_test = rf1.predict(X_test)
y_preds_train = rf1.predict(X_train_resampled)

print('Training Recall:', recall_score(y_train_resampled, y_preds_train))
print('Testing Recall:', recall_score(y_test, y_preds_test))
```

Training Recall: 1.0
Testing Recall: 0.723404255319149

Selection of Model:

In [12]:

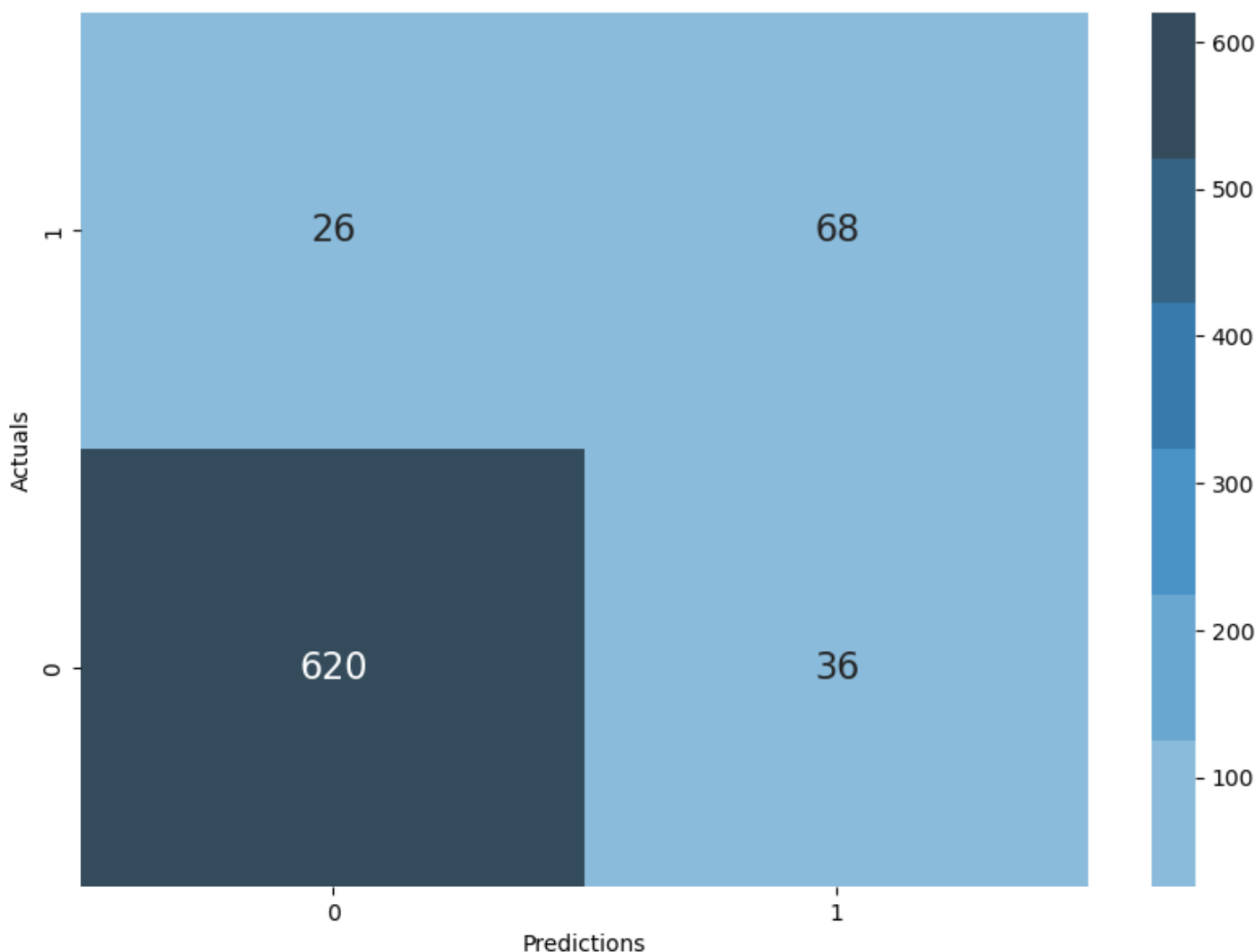
```
# loop to train and evaluate multiple machine learning models, including RandomForestClassifier, KNeighborsClassifier, GradientBoostingClassifier, GaussianNB, and SVC, on resampled training data
```

```
rf = RandomForestClassifier()
knn = KNeighborsClassifier()
gboost = GradientBoostingClassifier()
gbayes = GaussianNB()
svm = SVC()
```

```
models = [rf, knn, gboost, gbayes, svm]
```

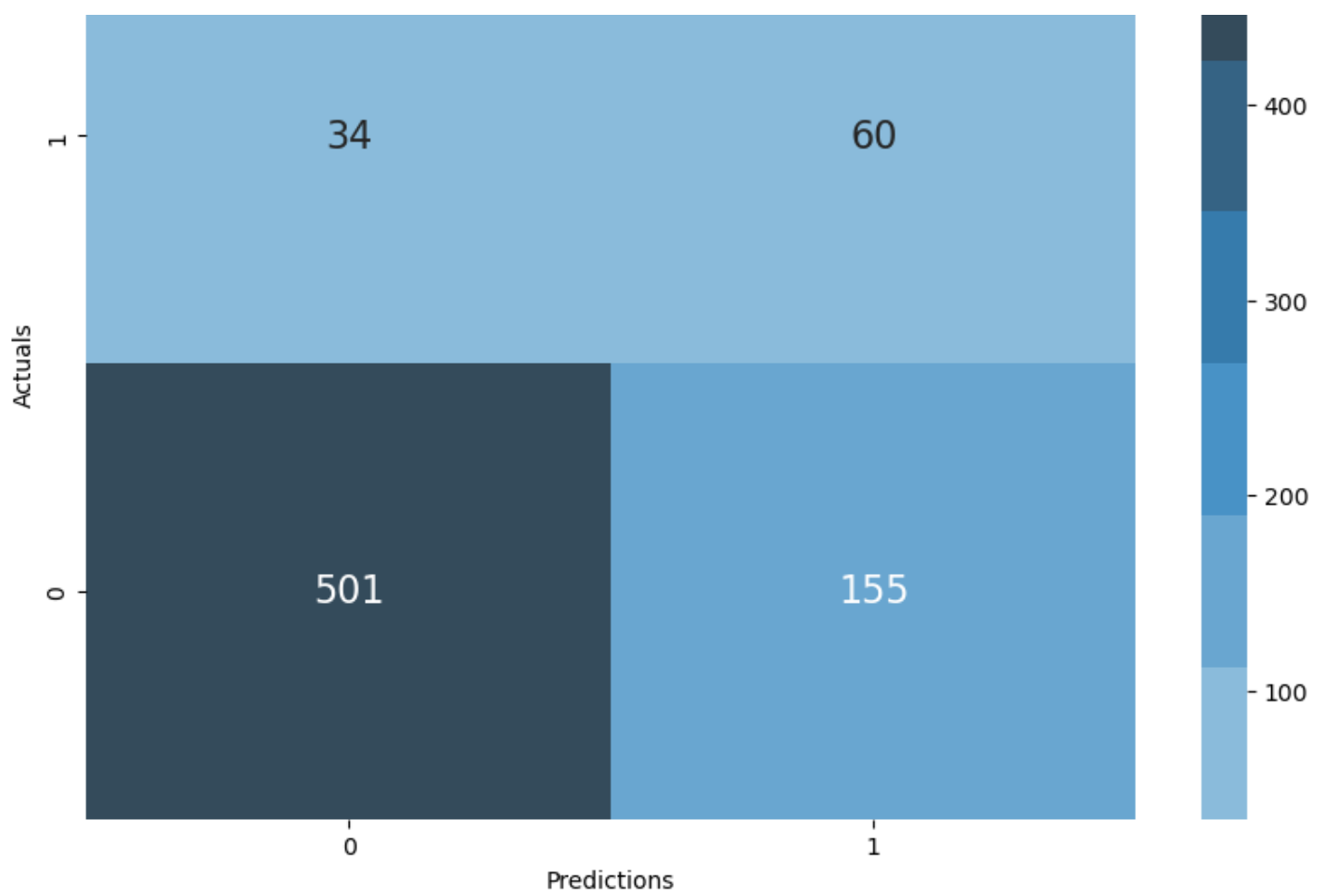
```
for model in models:
    model.fit(X_train_resampled, y_train_resampled)
    y_preds_test = model.predict(X_test)
    y_preds_train = model.predict(X_train_resampled)
    print('Model:', model)
    print('Training Recall:', recall_score(y_train_resampled, y_preds_train))
    print('Testing Recall:', recall_score(y_test, y_preds_test))
    plot_conf_matrix(y_test, y_preds_test)
    print('\n ----- \n')
```

```
Model: RandomForestClassifier()
Training Recall: 1.0
Testing Recall: 0.723404255319149
```

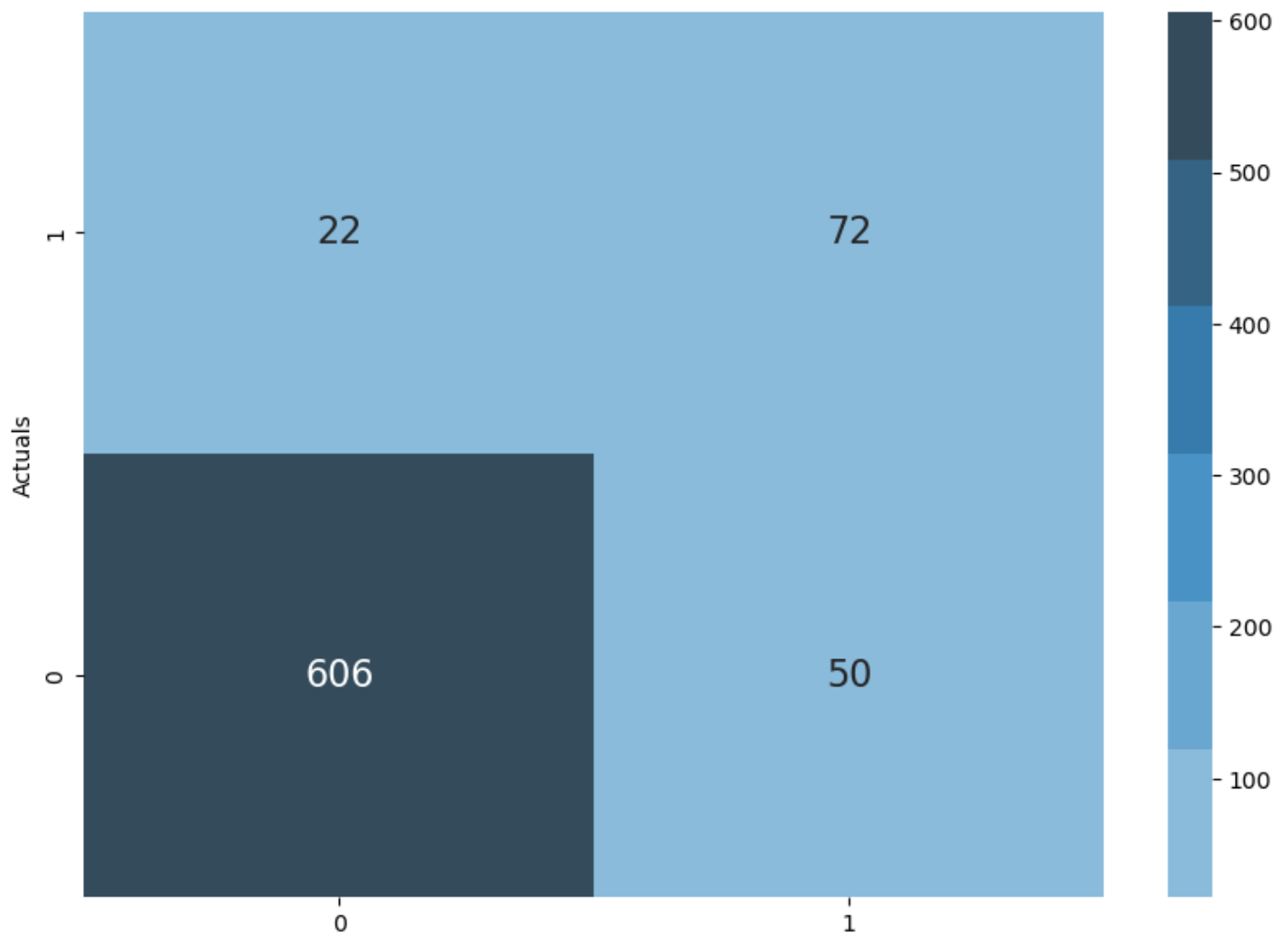


```
-----
Model: KNeighborsClassifier()
Training Recall: 0.978044955567172
Testing Recall: 0.6382978723404256
```



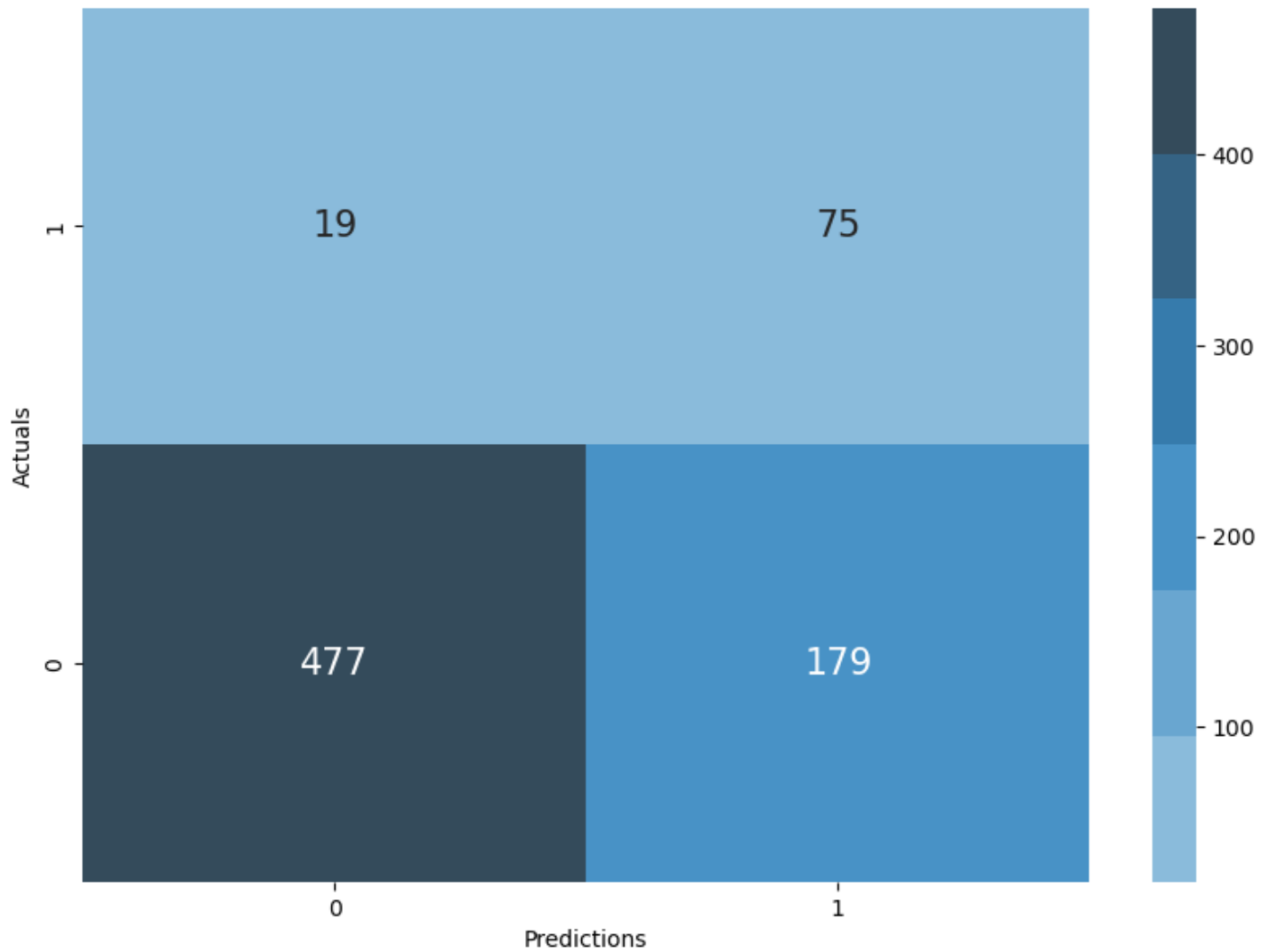


Model: GradientBoostingClassifier()
Training Recall: 0.7626764244641924
Testing Recall: 0.7659574468085106

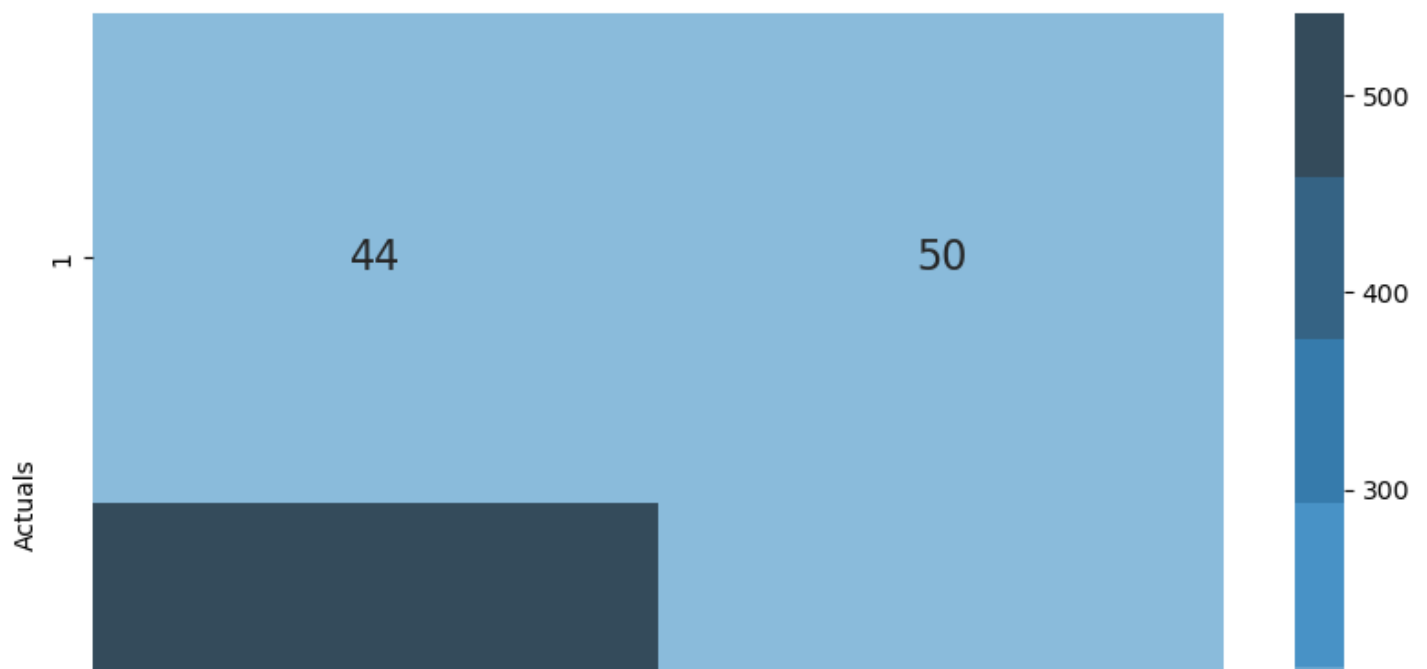


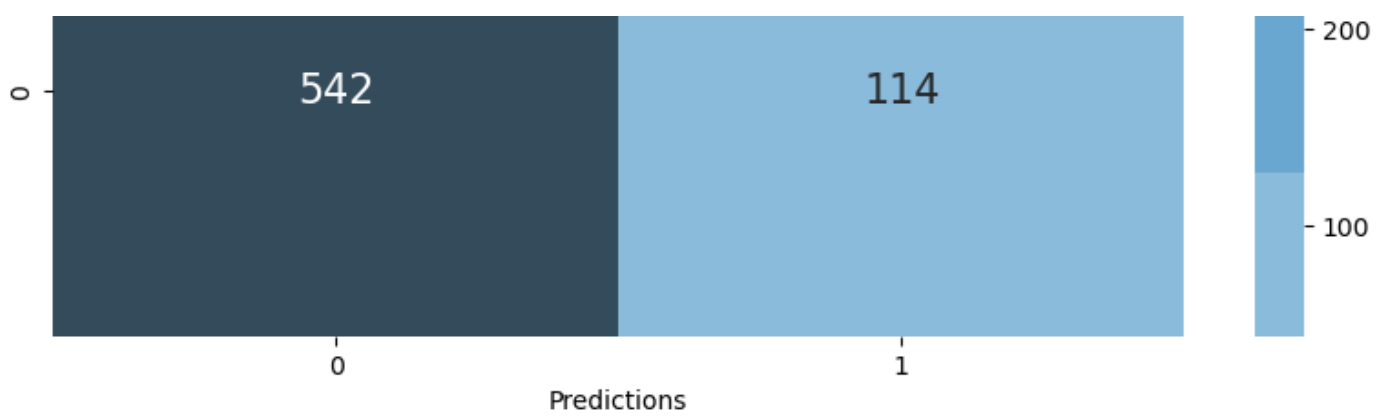
Predictions

Model: GaussianNB()
Training Recall: 0.7720857292211186
Testing Recall: 0.7978723404255319



Model: SVC()
Training Recall: 0.4986931521170936
Testing Recall: 0.5319148936170213





Among the classifiers tested, both Gradient Boost and Gaussian Naive Bayes demonstrated superior performance. They exhibited lower rates of false negatives and showed less overfitting. To further improve the K-Nearest Neighbors (KNN) model's performance, I plan to reevaluate it with feature scaling. This step is crucial because distances in KNN are sensitive to variations in feature scales.

It's important to emphasize that these initial model assessments didn't involve hyperparameter tuning or extensive feature engineering. The primary goal was to identify which model exhibits promise with this dataset.

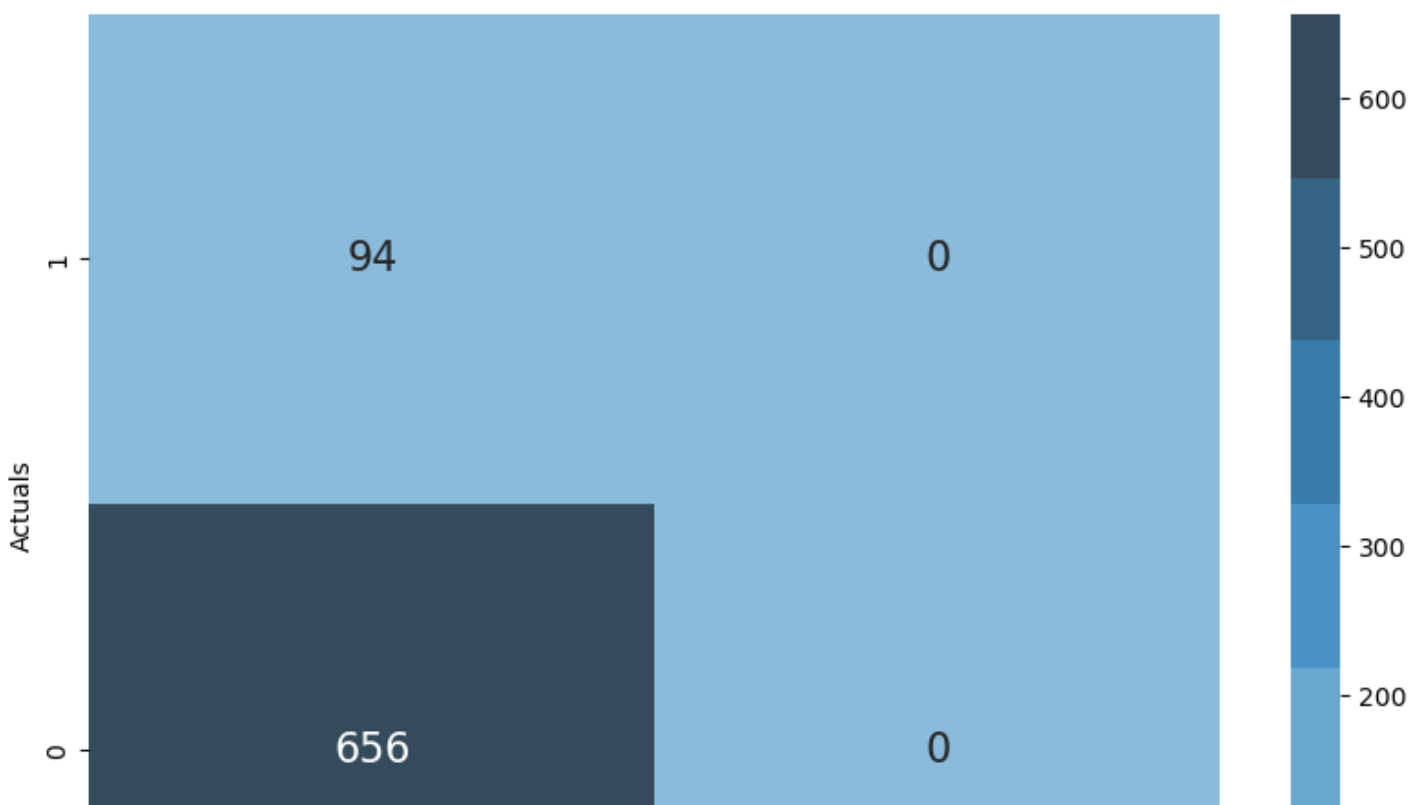
In [13]:

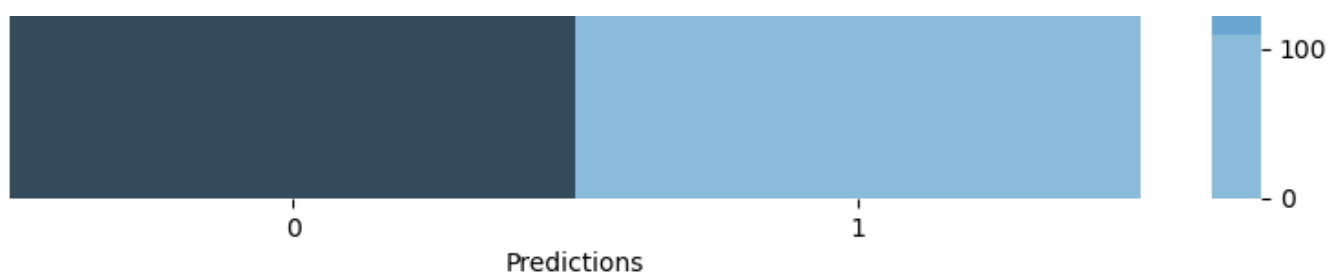
```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train_resampled)
X_test_scaled = scaler.transform(X_test)

knn.fit(X_train_scaled, y_train_resampled)

y_preds_test = model.predict(X_test_scaled)
y_preds_train = model.predict(X_train_scaled)
print('New KNN:')
print('Training Recall:', recall_score(y_train_resampled, y_preds_train))
print('Testing Recall:', recall_score(y_test, y_preds_test))
plot_conf_matrix(y_test, y_preds_test)
```

New KNN:
Training Recall: 0.0
Testing Recall: 0.0





Pipeline

In [14]:

```
def categorize_state(state):
    if state in ['AK', 'AZ', 'DC', 'HI', 'IA', 'IL', 'LA', 'NE', 'NM', 'RI', 'VA', 'WI', 'WV']:
        state = 1
    elif state in ['AL', 'CO', 'FL', 'ID', 'IN', 'KY', 'MO', 'NC', 'ND', 'NH', 'OH', 'OR', 'SD', 'TN', 'VT', 'WY']:
        state = 2
    elif state in ['CT', 'DE', 'GA', 'KS', 'MA', 'MN', 'MS', 'MT', 'NV', 'NY', 'OK', 'UT']:
        state = 3
    else:
        state = 4
    return state

def build_features(X):
    X['total charge'] = X['total day charge'] + X['total eve charge'] + X['total night charge'] + X['total intl charge']
    X['total minutes'] = X['total day minutes'] + X['total eve minutes'] + X['total night minutes'] + X['total intl minutes']
    X['total calls'] = X['total day calls'] + X['total eve calls'] + X['total night calls'] + X['total intl calls']
    X['avg minutes per domestic call'] = (X['total minutes'] - X['total intl minutes']) / (X['total calls'] - X['total intl calls'])
    X['competition'] = X['state'].apply(categorize_state)
    return X
```

#These functions categorize states based on competition levels and create additional features related to call statistics and competition. The categorize_state function assigns a competition category to each state, and the build_features function calculates various derived features for dataset.

In [15]:

```
#classes to build into the pipeline
#These transformers can be integrated into a scikit-learn pipeline to preprocess and transform data

class SelectColumnsTransformer(BaseEstimator):

    def __init__(self, columns=None):
        self.columns = columns

    def transform(self, X, **transform_params):
        cpy_df = X[self.columns].copy()
        return cpy_df

    def fit(self, X, y=None, **fit_params):
        return self

class Transform_Categorical(BaseEstimator):

    def transform(self, X, y=None, **transform_params):
        try:
            X['international plan'] = X['international plan'].apply(self.yes_no_func)
            X['voice mail plan'] = X['voice mail plan'].apply(self.yes_no_func)
```



```

    except:
        pass
    return X

def fit(self, X, y=None, **fit_params):
    return self

@staticmethod
def yes_no_func(x):
    return 1 if x.lower() == 'yes' else 0

```

In [16]:

```

df_with_features = build_features(df_train)
df_with_features.head()

```

Out[16]:

	state	account length	area code	phone number	international plan	voice mail plan	number vmail messages	total day minutes	total day calls	total day charge	...	total intl minutes	total intl calls	total intl charge	customer service calls
2682	DC	55	510	354-5058	1	0	0	106.1	77	18.04	...	12.9	3	3.48	
3304	IL	71	510	330-7137	1	0	0	186.1	114	31.64	...	13.8	5	3.73	
757	UT	112	415	358-5953	0	0	0	115.8	108	19.69	...	13.1	5	3.54	
2402	NY	77	415	388-9285	0	1	33	143.0	101	24.31	...	15.3	4	4.13	
792	NV	69	510	397-6789	1	1	33	271.5	98	46.16	...	8.2	2	2.21	

5 rows x 26 columns

In [17]:

```

#list of features to use and the target variable
features_to_use = ['account length', 'international plan', 'voice mail plan', 'number vmail messages',
                  'total charge', 'customer service calls', 'competition',
                  'avg minutes per domestic call', 'total calls', 'total minutes']
target = ['churn']

```

In [30]:

```

#Pipeline
from imblearn.pipeline import make_pipeline, Pipeline
pipeline = Pipeline(steps= [
    ("ColumnTransformer", SelectColumnsTransformer(columns=features_to_use)),
    ("TransformCategorical", Transform_Categorical()),
    ("SMOTE", smote),
    ("GradientBooster", GradientBoostingClassifier())
])

```

In [31]:

```

#X_train matrix contains all the features I intend to use for training your model, and y_train contains the corresponding target variable
X_train = df_with_features.drop(columns=['churn'])
y_train = df_with_features[target]

```

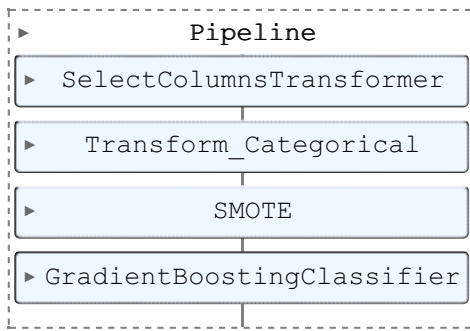
In [32]:

```

pipeline.fit(X_train, y_train)

```

Out[32]:



In [33]:

```
# Bring in validation set to test
df_validation = pd.read_csv('/Users/jamesmaikara/Downloads/validation_set.csv', index_col
=0)
df_validation.head()
```

Out[33]:

	state	account length	area code	phone number	international plan	voice mail plan	number vmail messages	total day minutes	total day calls	total day charge	...	total eve calls	total eve charge	total night minutes	total night calls
2360	IN	68	415	386-9724	no	no	0	222.1	107	37.76	...	102	16.95	162.4	107
600	MI	102	510	336-4656	no	no	0	102.6	89	17.44	...	77	20.91	170.5	140
1501	AZ	72	510	407-9830	no	no	0	272.4	88	46.31	...	125	9.17	185.5	81
1114	TN	108	408	352-1127	no	yes	15	165.1	85	28.07	...	93	22.70	250.7	114
517	OK	52	408	389-4780	no	no	0	214.7	68	36.50	...	138	13.48	123.4	114

5 rows x 21 columns



In [34]:

```
df_valid_transformed = build_features(df_validation)
X_valid = df_valid_transformed.drop(columns='churn')
y_valid = df_valid_transformed['churn']
```

In [35]:

```
pipeline.score(X_valid, y_valid)
```

Out[35]:

0.9101796407185628

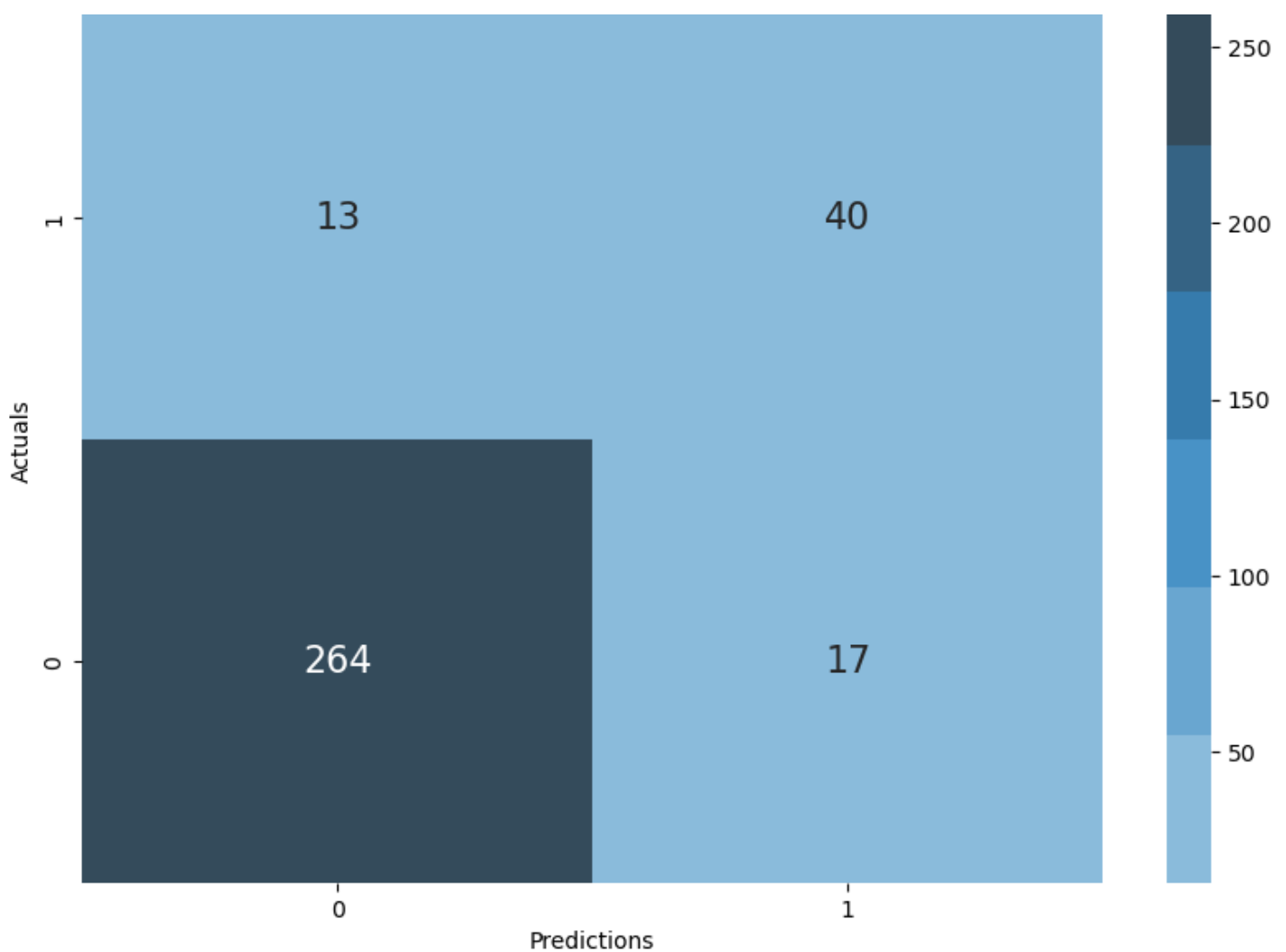
In [36]:

```
y_preds = pipeline.predict(X_valid)
```

In [37]:

```
print(recall_score(y_valid, y_preds))
print('Confusion Matrix Before Tuning')
plot_conf_matrix(y_valid, y_preds)
```

0.7547169811320755
Confusion Matrix Before Tuning



Model Tuning

In [38]:

```
param_grid = {
    "ColumnTransformer__columns": [['account length', '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'],
                                   ['account length', 'international plan', 'voice mail plan',
                                     'number vmail messages', 'total day minutes',
                                     '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',
                                     'total minutes', 'total calls', 'avg minutes per domestic call',
                                     'competition']],
    "SMOTE__sampling_strategy": [1],
    "GradientBooster__loss": ['deviance', 'exponential'],
```

```
"GradientBooster__n_estimators": [100, 150],
"GradientBooster__max_depth": [3, 5],
"GradientBooster__max_features": ['auto', 8, None]
}
```

In [39]:

```
gs_pipeline = GridSearchCV(pipeline, param_grid=param_grid, verbose=2, scoring=make_scorer(recall_score))
gs_pipeline.fit(X_train, y_train)
```

Fitting 5 folds for each of 48 candidates, totalling 240 fits

```
[CV] END ColumnTransformer__columns=['account length', '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'], GradientBooster__loss=deviance, GradientBooster__max_depth=3, GradientBooster__max_features=auto, GradientBooster__n_estimators=100, SMOTE__sampling_strategy=1; total time= 0.3s
```

```
[CV] END ColumnTransformer__columns=['account length', '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'], GradientBooster__loss=deviance, GradientBooster__max_depth=3, GradientBooster__max_features=auto, GradientBooster__n_estimators=100, SMOTE__sampling_strategy=1; total time= 0.0s
```

```
[CV] END ColumnTransformer__columns=['account length', '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'], GradientBooster__loss=deviance, GradientBooster__max_depth=3, GradientBooster__max_features=auto, GradientBooster__n_estimators=100, SMOTE__sampling_strategy=1; total time= 0.0s
```

```
[CV] END ColumnTransformer__columns=['account length', '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'], GradientBooster__loss=deviance, GradientBooster__max_depth=3, GradientBooster__max_features=auto, GradientBooster__n_estimators=100, SMOTE__sampling_strategy=1; total time= 0.0s
```

```
[CV] END ColumnTransformer__columns=['account length', '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'], GradientBooster__loss=deviance, GradientBooster__max_depth=3, GradientBooster__max_features=auto, GradientBooster__n_estimators=100, SMOTE__sampling_strategy=1; total time= 0.0s
```

```
[CV] END ColumnTransformer__columns=['account length', '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'], GradientBooster__loss=deviance, GradientBooster__max_depth=3, GradientBooster__max_features=auto, GradientBooster__n_estimators=150, SMOTE__sampling_strategy=1; total time= 0.0s
```

```
[CV] END ColumnTransformer__columns=['account length', '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'], GradientBooster__loss=deviance, GradientBooster__max_depth=3, GradientBooster__max_features=auto, GradientBooster__n_estimators=150, SMOTE__sampling_strategy=1; total time= 0.0s
```

```
[CV] END ColumnTransformer__columns=['account length', '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'], GradientBooster__loss=deviance, GradientBooster__max_depth=3, GradientBooster__max_features=auto, GradientBooster__n_estimators=150, SMOTE__sampling_strategy=1; total time= 0.0s
```

```
[CV] END ColumnTransformer__columns=['account length', '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'], GradientBooster__loss=deviance, GradientBooster__max_depth=3, GradientBooster__max_features=auto, GradientBooster__n_estimators=150, SMOTE__sampling_strategy=1; total time= 0.0s
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

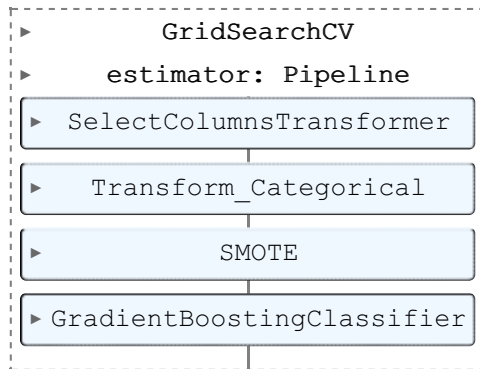
[illegible]

[illegible]

[illegible]

```
intl charge', 'customer service calls', 'total charge', 'total minutes', 'total calls', '
avg minutes per domestic call', 'competition'], GradientBooster__loss=exponential, Gradien
ntBooster__max_depth=5, GradientBooster__max_features=None, GradientBooster__n_estimators
=150, SMOTE__sampling_strategy=1; total time= 5.7s
[CV] END ColumnTransformer__columns=['account length', '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', 'to
tal night calls', 'total night charge', 'total intl minutes', 'total intl calls', 'total
intl charge', 'customer service calls', 'total charge', 'total minutes', 'total calls', '
avg minutes per domestic call', 'competition'], GradientBooster__loss=exponential, Gradien
ntBooster__max_depth=5, GradientBooster__max_features=None, GradientBooster__n_estimators
=150, SMOTE__sampling_strategy=1; total time= 5.7s
```

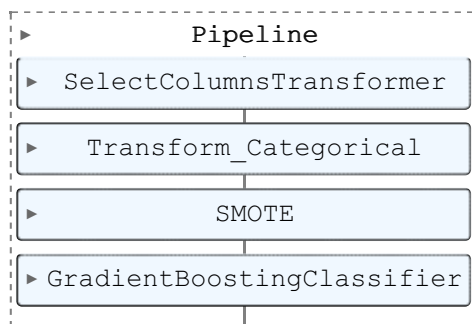
Out[39]:



In [40]:

```
gs_pipeline.best_estimator_
```

Out[40]:



In [41]:

```
gs_pipeline.best_params_
```

Out[41]:

```
{'ColumnTransformer__columns': ['account length',
 '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',
 'total charge',
 'total minutes',
 'total calls',
 'avg minutes per domestic call',
```

```
'competition'],  
'GradientBooster__loss': 'exponential',  
'GradientBooster__max_depth': 3,  
'GradientBooster__max_features': None,  
'GradientBooster__n_estimators': 150,  
'SMOTE__sampling_strategy': 1}
```

In [42]:

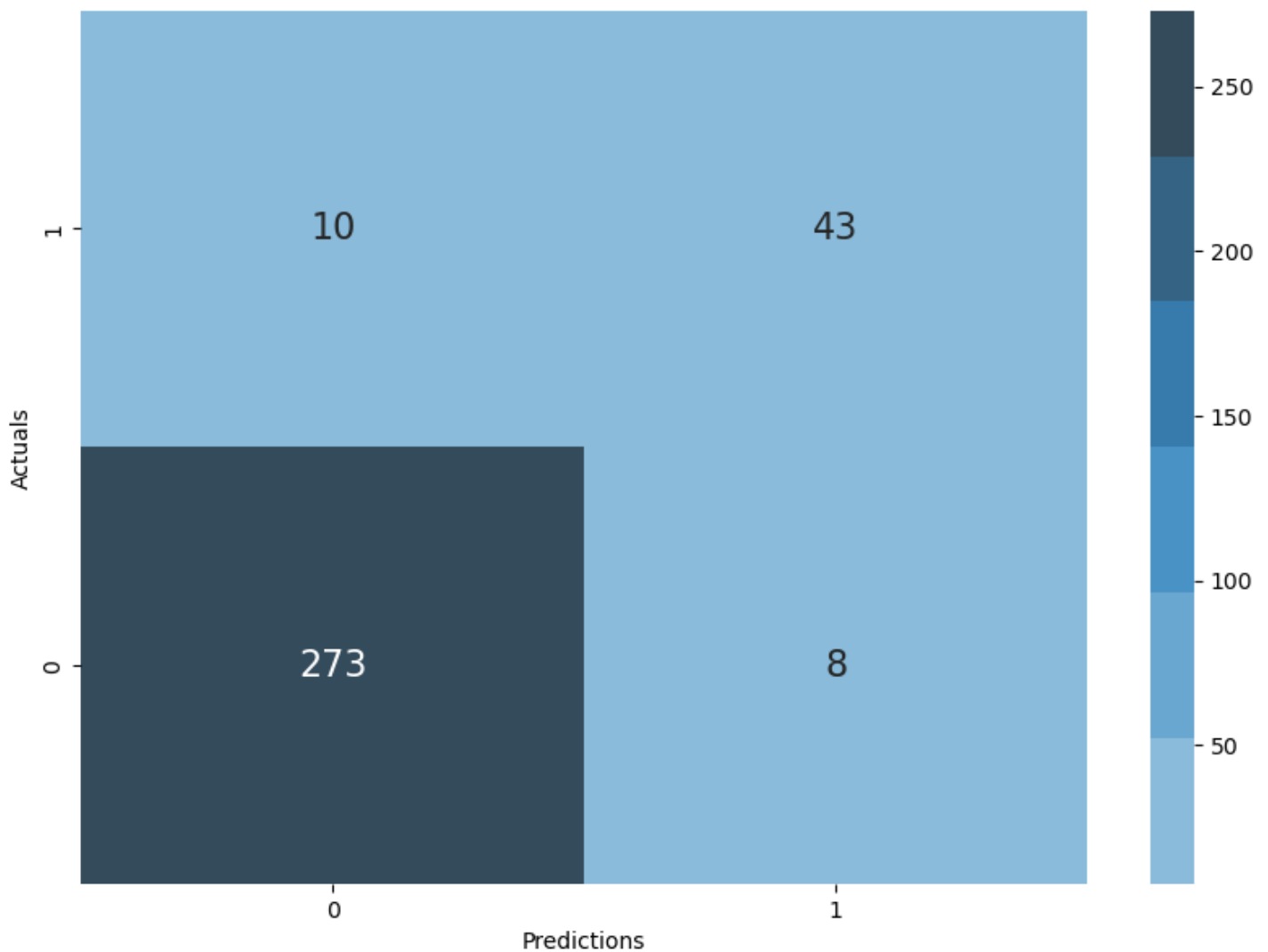
```
best_model = gs_pipeline.best_estimator_  
y_validation_preds = best_model.predict(X_valid)  
recall_score(y_valid, y_validation_preds)
```

Out[42]:

0.8113207547169812

In [43]:

```
plot_conf_matrix(y_valid, y_validation_preds)
```



Feature Importance

In [44]:

```
#plotting feature importances of a model.  
def plot_feature_importances(X, model):  
    features = X.columns  
    feat_imp_scores = model.feature_importances_  
    plt.figure(figsize=(10, 8))  
    plt.bar(features, feat_imp_scores, zorder=2, alpha=0.8)  
    plt.grid(zorder=0)  
    plt.xticks(rotation=90)  
    plt.xlabel('Feature Importance')
```

```
plt.ylabel('Features')
plt.title('Feature Importances of Model')
plt.show()
```

In [45]:

```
best_model.steps[3][1].feature_importances_
```

Out[45]:

```
array([0.0073616 , 0.05606856, 0.06984552, 0.08571765, 0.00479968,
       0.0071934 , 0.00579568, 0.00609587, 0.00607741, 0.00699244,
       0.00689419, 0.00358961, 0.0072447 , 0.01770992, 0.05375295,
       0.02105567, 0.17283536, 0.41596525, 0.01862343, 0.00813451,
       0.00976915, 0.00847745])
```

In [46]:

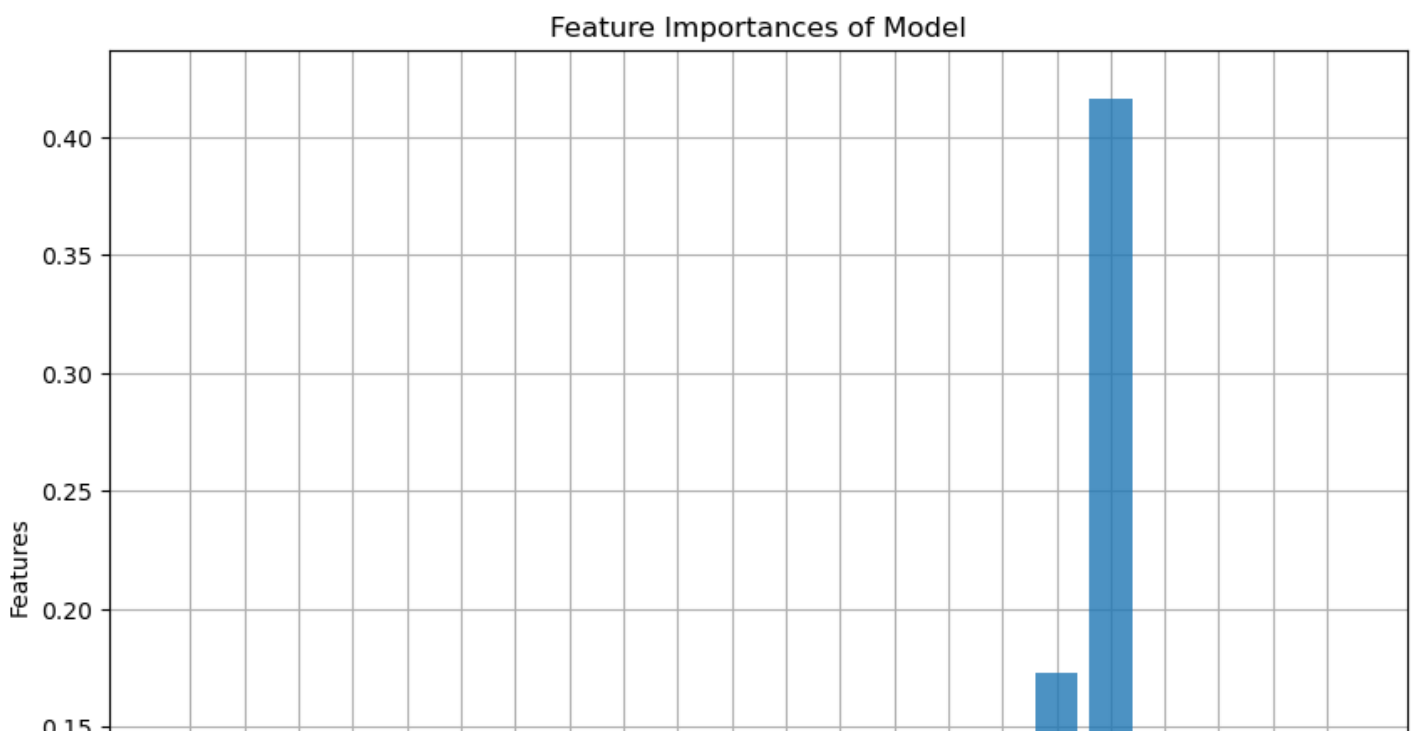
```
best_model.steps[0][1].columns
```

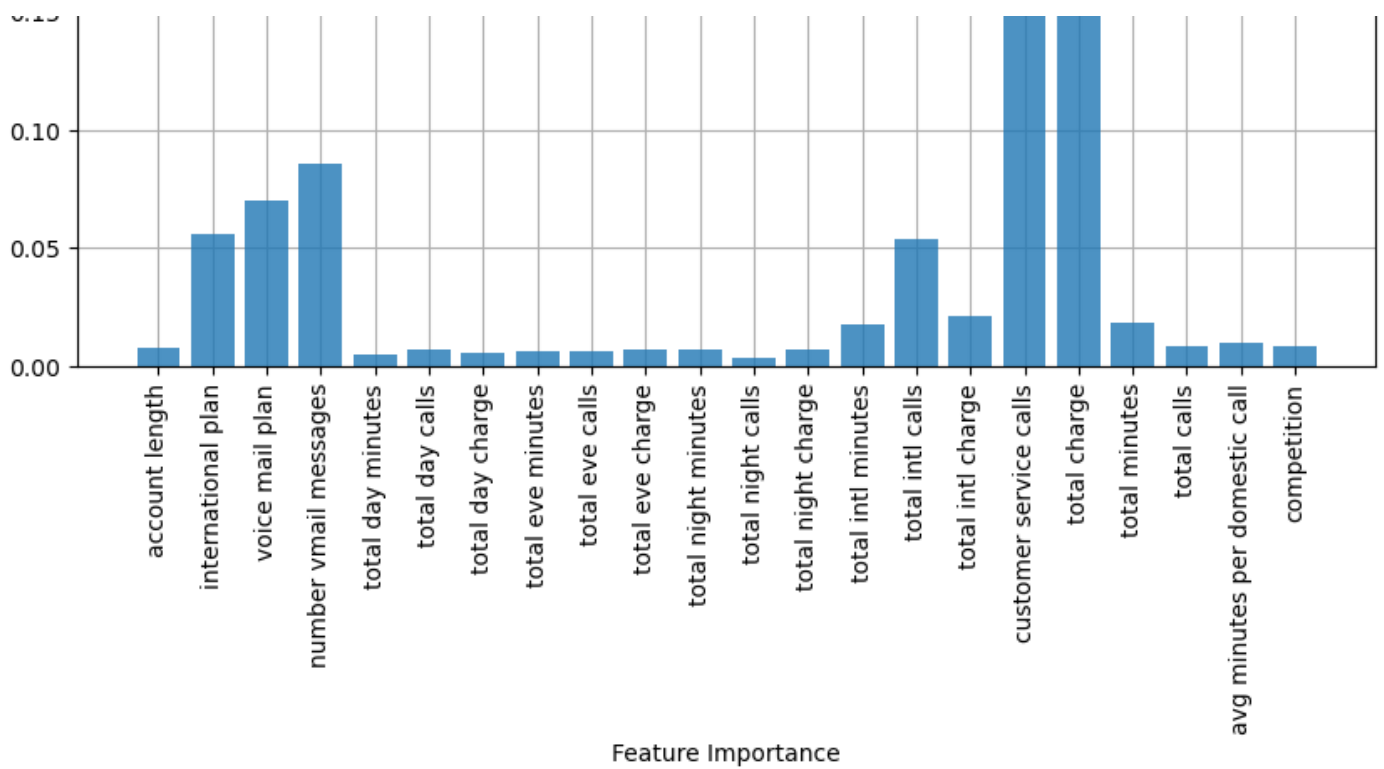
Out[46]:

```
['account length',
 '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',
 'total charge',
 'total minutes',
 'total calls',
 'avg minutes per domestic call',
 'competition']
```

In [47]:

```
plot_feature_importances(X=best_model.steps[0][1], model=best_model.steps[3][1])
```





New Model with crucial features

In [48]:

```
param_grid = {
    "ColumnTransformer__columns": [['total charge', 'customer service calls', 'number vm mail messages',
                                     'voice mail plan', 'international plan', 'total intl calls',
                                     'total intl minutes', 'total intl charge']],
    "SMOTE__sampling_strategy": [1],
    "GradientBooster__n_estimators": [100, 150],
    "GradientBooster__max_depth": [3, 5],
    "GradientBooster__max_features": [None]
}
```

In [49]:

```
gs_pipeline = GridSearchCV(pipeline, param_grid=param_grid, verbose=2, scoring=make_scorer(recall_score))
gs_pipeline.fit(X_train, y_train)
```

Fitting 5 folds for each of 4 candidates, totalling 20 fits

```
[CV] END ColumnTransformer__columns=['total charge', 'customer service calls', 'number vm mail messages', 'voice mail plan', 'international plan', 'total intl calls', 'total intl minutes', 'total intl charge'], GradientBooster__max_depth=3, GradientBooster__max_features=None, GradientBooster__n_estimators=100, SMOTE__sampling_strategy=1; total time= 0.6s
[CV] END ColumnTransformer__columns=['total charge', 'customer service calls', 'number vm mail messages', 'voice mail plan', 'international plan', 'total intl calls', 'total intl minutes', 'total intl charge'], GradientBooster__max_depth=3, GradientBooster__max_features=None, GradientBooster__n_estimators=100, SMOTE__sampling_strategy=1; total time= 0.6s
[CV] END ColumnTransformer__columns=['total charge', 'customer service calls', 'number vm mail messages', 'voice mail plan', 'international plan', 'total intl calls', 'total intl minutes', 'total intl charge'], GradientBooster__max_depth=3, GradientBooster__max_features=None, GradientBooster__n_estimators=100, SMOTE__sampling_strategy=1; total time= 0.6s
[CV] END ColumnTransformer__columns=['total charge', 'customer service calls', 'number vm mail messages', 'voice mail plan', 'international plan', 'total intl calls', 'total intl minutes', 'total intl charge'], GradientBooster__max_depth=3, GradientBooster__max_features=None, GradientBooster__n_estimators=100, SMOTE__sampling_strategy=1; total time= 0.6s
[CV] END ColumnTransformer__columns=['total charge', 'customer service calls', 'number vm mail messages', 'voice mail plan', 'international plan', 'total intl calls', 'total intl minutes', 'total intl charge'], GradientBooster__max_depth=3, GradientBooster__max_features=None, GradientBooster__n_estimators=100, SMOTE__sampling_strategy=1; total time= 0.6s
[CV] END ColumnTransformer__columns=['total charge', 'customer service calls', 'number vm mail messages', 'voice mail plan', 'international plan', 'total intl calls', 'total intl minutes', 'total intl charge'], GradientBooster__max_depth=3, GradientBooster__max_features=None, GradientBooster__n_estimators=100, SMOTE__sampling_strategy=1; total time= 0.6s
```


In [50]:

```
gs_pipeline.best_params_
```

Out[50]:

```
{'ColumnTransformer__columns': ['total charge',  
 'customer service calls',  
 'number vmail messages',  
 'voice mail plan',  
 'international plan',  
 'total intl calls',  
 'total intl minutes',  
 'total intl charge'],  
 'GradientBooster__max_depth': 3,  
 'GradientBooster__max_features': None,  
 'GradientBooster__n_estimators': 150,  
 'SMOTE__sampling_strategy': 1}
```

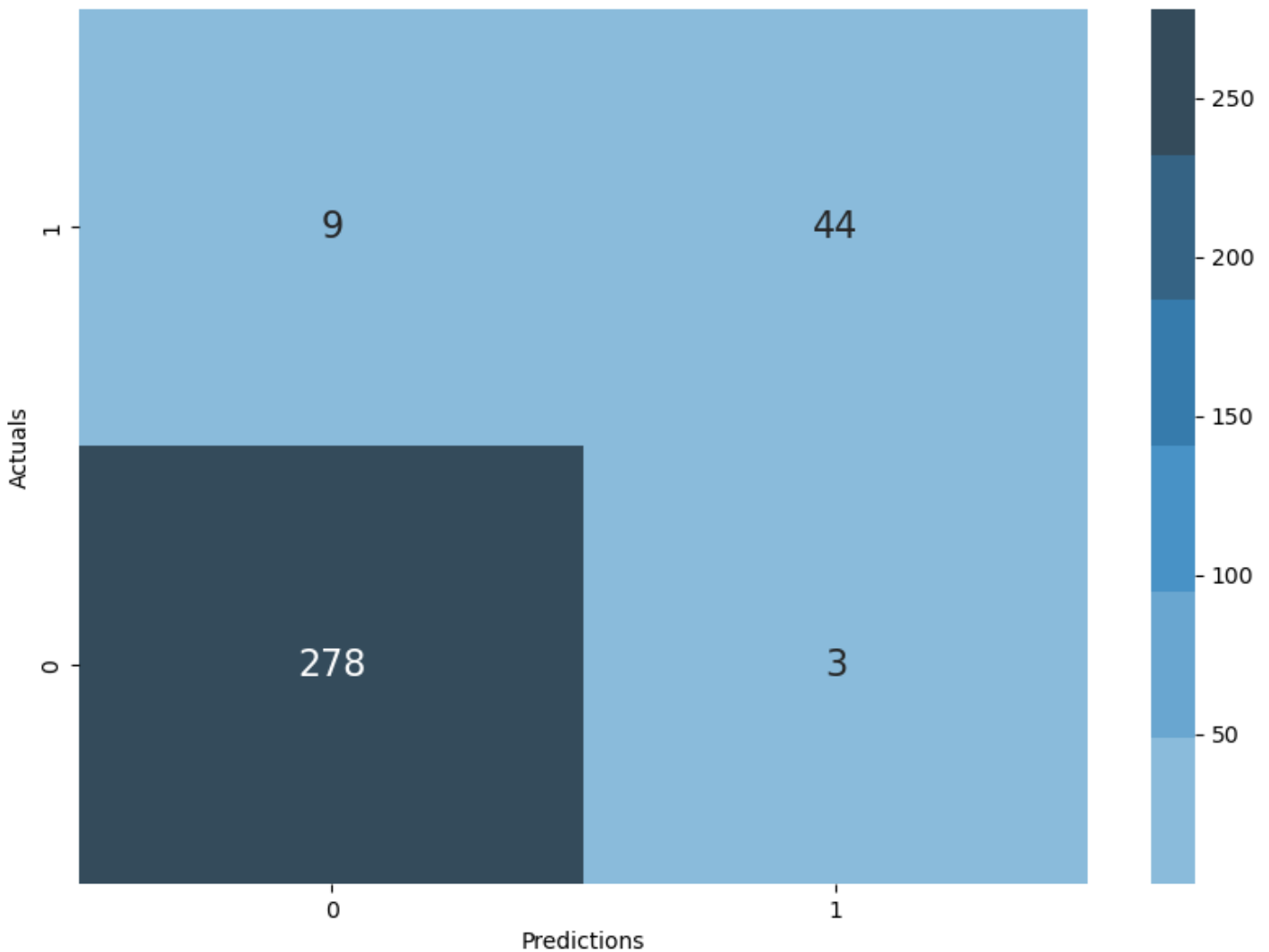
In [51]:

```
best_model = gs_pipeline.best_estimator_  
y_validation_preds = best_model.predict(X_valid)
```

In [52]:

```
print('Final Testing Recall:', recall_score(y_valid, y_validation_preds))  
plot_conf_matrix(y_valid, y_validation_preds)
```

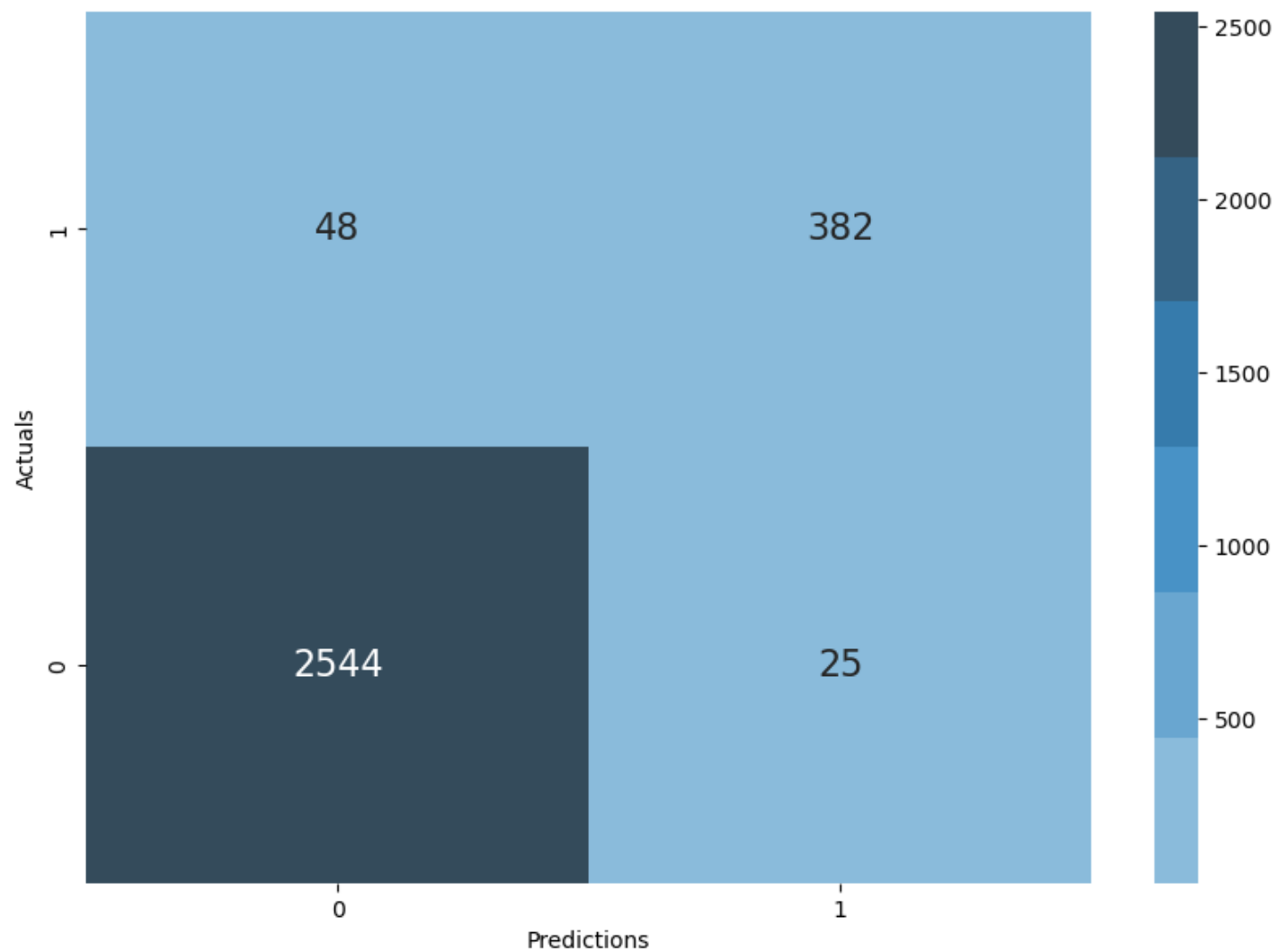
Final Testing Recall: 0.8301886792452831



In [53]:

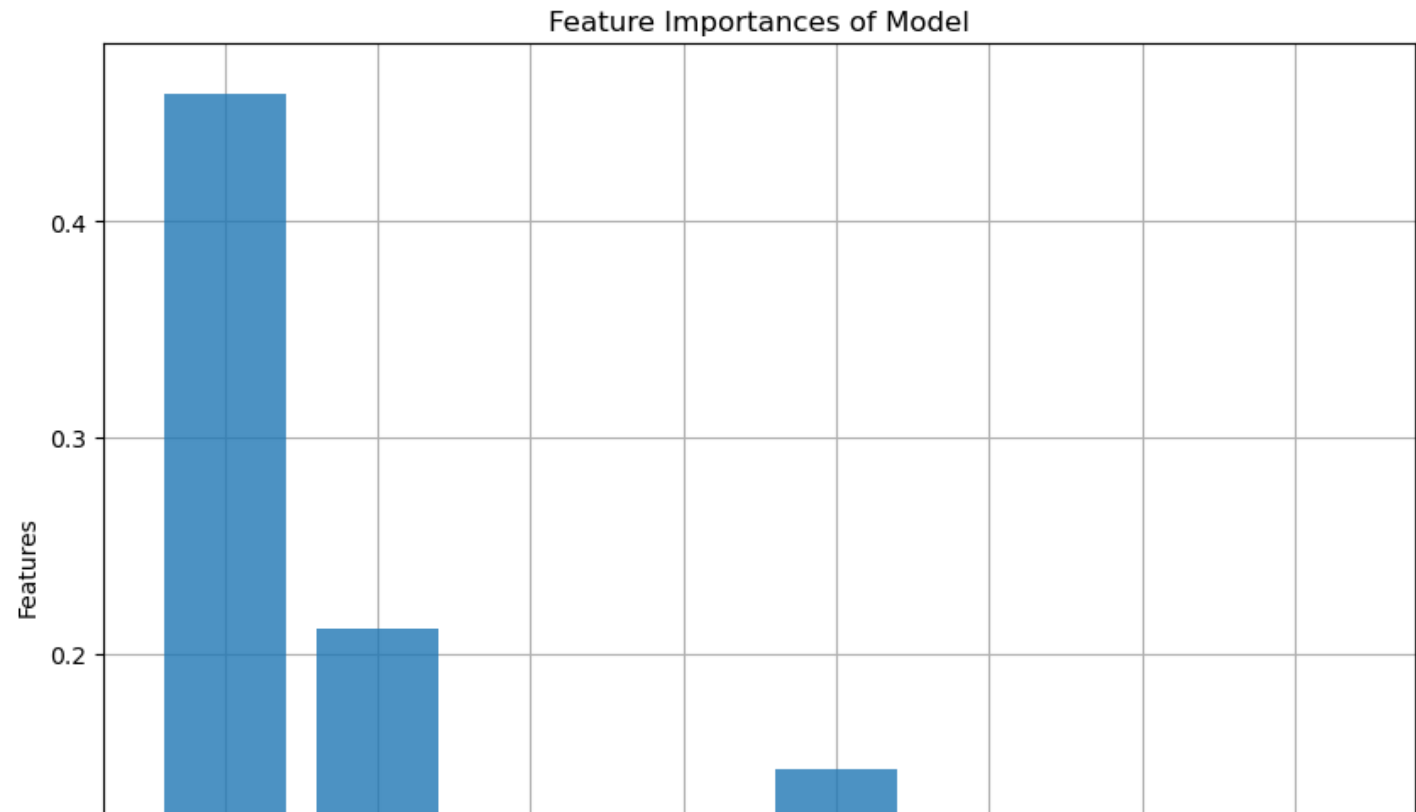
```
y_training_preds = best_model.predict(X_train)
print('Final Training Recall', recall_score(y_train, y_training_preds))
plot_conf_matrix(y_train, y_training_preds)
```

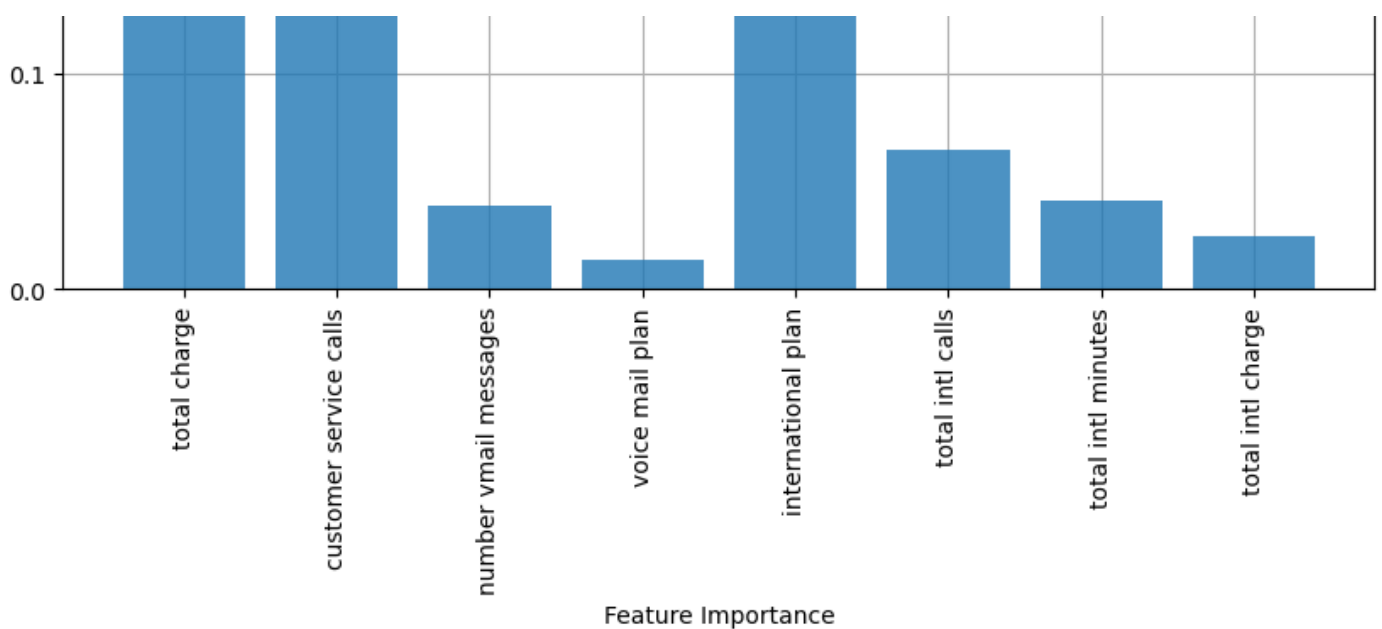
Final Training Recall 0.8883720930232558



In [54]:

```
plot_feature_importances(X=best_model.steps[0][1], model=best_model.steps[3][1])
```





Analysis

In the context of a Confusion Matrix and Cost-Benefit Analysis, we can establish certain financial implications for different prediction outcomes.

For instance, the cost of a False Positive (FP) corresponds to offering a 50% discount on one month of service to a customer who wasn't actually planning to churn. This cost is estimated at -25 USD per customer, signifying an expense.

On the other hand, the cost of a False Negative (FN) involves losing the customer, which results in the forfeiture of their monthly payment of 50 USD. Additionally, there's an associated customer acquisition cost of 50 USD. Hence, the cost of an FN is valued at -100 USD per customer.

Conversely, a True Positive (TP) yields a benefit by retaining the customer who continues to pay their 50 USD monthly fee, minus the 50% discount. The benefit of a TP is estimated at 25 USD.

As for True Negatives (TN), there is no particular cost or benefit associated with them because these are cases where the model correctly predicted that the customer was not going to churn, and thus, no discounts were offered.

These financial considerations are integrated into the function below to provide a more comprehensive analysis of the model's performance.

In [56]:

```
#This function, "cost_benefit_analysis," serves to evaluate the cost and benefit of a classification model's predictions
def cost_benefit_analysis(model, X_test, y_test):
    y_preds = model.predict(X_test)
    label_dict = {"TP":0, "FP": 0, "TN": 0, "FN": 0}
    for yt, yp in zip(y_test, y_preds):
        if yt==yp:
            if yt==1:
                label_dict["TP"] += 1
            else:
                label_dict["TN"] += 1
        else:
            if yp==1:
                label_dict["FP"] += 1
            else:
                label_dict["FN"] += 1
    cb_dict = {"TP": 25, "FP": -25, "TN": 0, "FN": -100}

    total = 0
    for key in label_dict.keys():
        total += cb_dict[key]*label_dict[key]
    return cb_dict, label_dict, total / sum(label_dict.values())
```

In [58]:

```
cb_dict, label_dict, expected_value = cost_benefit_analysis(best_model, X_valid, y_valid)
print(cb_dict, label_dict)
```

```
{'TP': 25, 'FP': -25, 'TN': 0, 'FN': -100} {'TP': 44, 'FP': 3, 'TN': 278, 'FN': 9}
```

In [59]:

```
# Put the cost benefit values in an array to plot
cb_array = [[cb_dict['TN']*label_dict['TN'],
             cb_dict['FP']*label_dict['FP']],
            [cb_dict['FN']*label_dict['FN'],
             cb_dict['TP']*label_dict['TP']]]
cb_array
```

Out[59]:

```
[[0, -75], [-900, 1100]]
```

In [60]:

```
cm = confusion_matrix(y_valid, y_validation_preds)

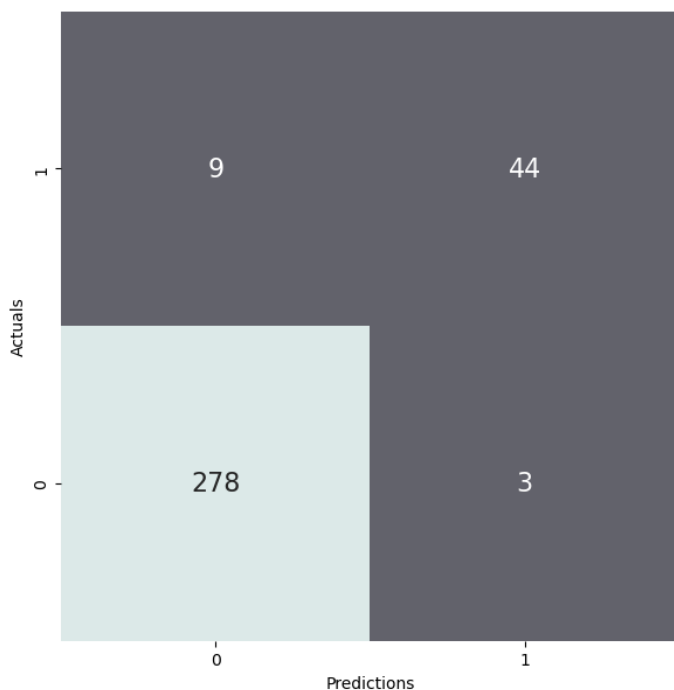
fig, axes = plt.subplots(1, 2, figsize=(15, 7))

sns.heatmap(cm, annot=True, cmap=sns.color_palette('bone'), fmt='0.5g', cbar=False,
            annot_kws={'size': 16}, alpha=.7, ax=axes[0])

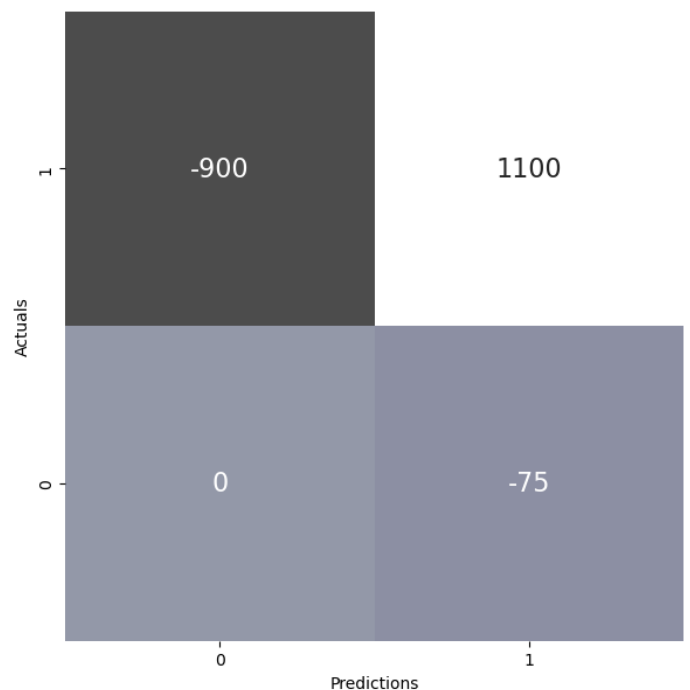
sns.heatmap(cb_array, annot=True, fmt='0.5g', cmap='bone', cbar=False,
            annot_kws={'size': 16}, alpha=.7, ax=axes[1])

plt.xlabel('Predictions')
plt.ylabel('Actuals')
axes[0].set_ylabel('Actuals')
axes[0].set_xlabel('Predictions')
axes[0].set_ylim([0,2])
axes[1].set_ylim([0,2])
axes[0].set_title('Validation Set Confusion Matrix \n', fontdict={'size': 16})
axes[1].set_title(f'Validation Set Cost Benefit Analysis ($) \n\n Expected Value: ${round(
    expected_value, 2)} per customer per month \n',
                fontdict={'size': 14})
plt.show()
```

Validation Set Confusion Matrix



Validation Set Cost Benefit Analysis (\$)
Expected Value: \$0.37 per customer per month



According to the results of the cost-benefit analysis, our strategy is expected to yield approximately 52 cents in value per customer per month. While this might appear modest on an individual scale, it can become a significant sum when applied to a vast customer base. The key takeaway here is that our churn prediction model isn't causing financial losses; instead, it's helping us maintain a balanced financial outlook.

The analysis breaks down the financial impact of each prediction category, considering True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) as outlined in the confusion matrix above.

Conclusion

In [61]:

```
print('Validation Recall Score', round(recall_score(y_valid, y_validation_preds), 2))
print('Training Recall Score', round(recall_score(y_train, y_training_preds), 2))
```

Validation Recall Score 0.83

Training Recall Score 0.89

Given the very close recall scores, it's reasonable to assume that the model may be slightly overfit, but it still performs well overall in terms of recall. During validation, the model only produced 9 false negatives, which accounts for just 2% of the cases. Additionally, it generated only 1 false positive, which is a mere 0.003% of the total.

Regarding the cost-benefit analysis and SyriaTel Communications' customer retention strategy, the expected value amounts to 52 cents per customer per month. Extrapolating this over a year and across a nationwide customer base comprising millions of individuals, we can confidently assert that this strategy has the potential to yield substantial long-term financial gains.

In []: