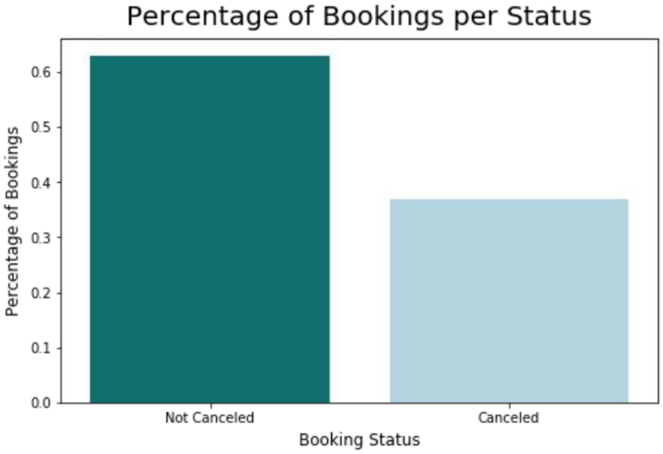
The goal of this project is to build a model able to classify a booking as canceled or not canceled. To achieve this goal, I started with Exploratory Data Analysis to gain insights about the customers and hopefully reasons why they cancel their reservation. Then I created a classification model to predict whether or not a booking will be canceled with the highest accuracy possible. This model will allow hotels to predict if a new booking will be canceled or not, manage their business accordingly, and increase their revenue.

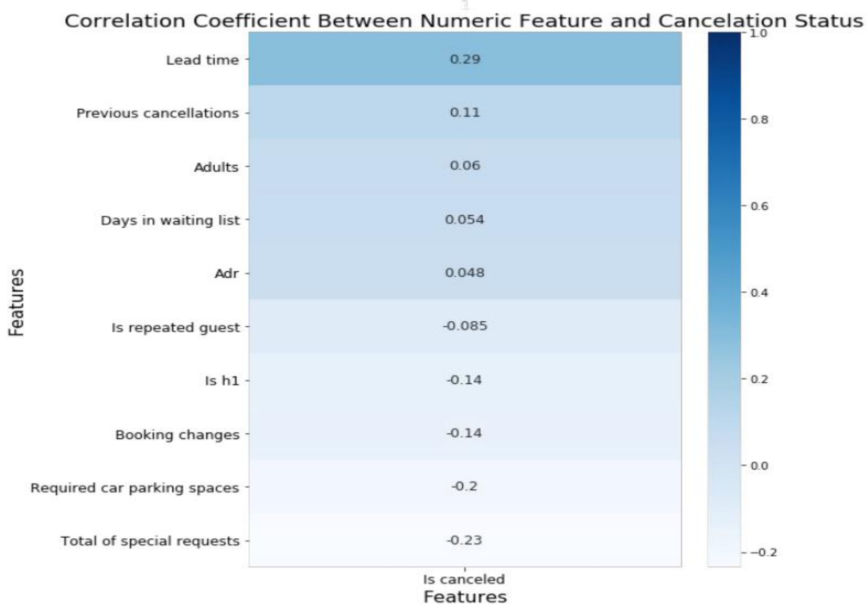
Exploratory Data Analysis

- Visualizing the percentage of canceled vs not canceled bookings:



As shown above 37% of bookings were canceled.

- Visualizing correlation coefficients between features and cancellation:

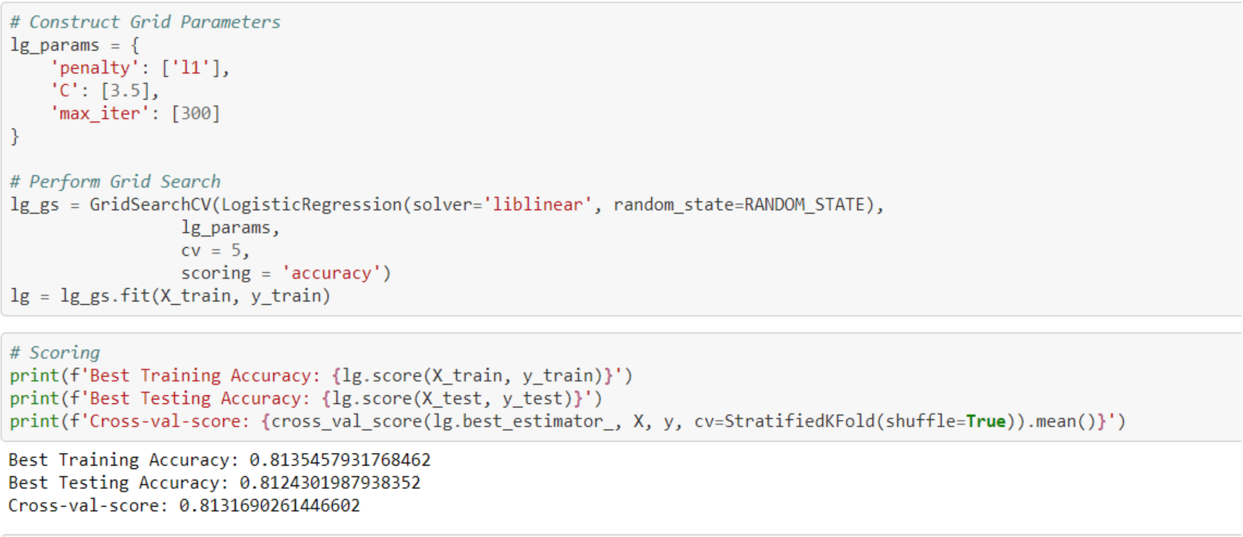


Based on above, I can conclude than the features who have the biggest correlation to the target ('is cancelled') are as follows:

a. **Lead Time**
b. **Special Requests**
c. **Parking Spaces**
d. **Booking Changes**
e. **Previous Cancellations**

Modeling

- Logistic Regression:



```
# Construct Grid Parameters
lg_params = {
    'penalty': ['l1'],
    'C': [3.5],
    'max_iter': [300]
}

# Perform Grid Search
lg_gs = GridSearchCV(LogisticRegression(solver='liblinear', random_state=RANDOM_STATE),
                     lg_params,
                     cv = 5,
                     scoring = 'accuracy')
lg = lg_gs.fit(X_train, y_train)
```

```
# Scoring
print(f'Best Training Accuracy: {lg.score(X_train, y_train)}')
print(f'Best Testing Accuracy: {lg.score(X_test, y_test)}')
print(f'Cross-val-score: {cross_val_score(lg.best_estimator_, X, y, cv=StratifiedKFold(shuffle=True)).mean()}')
```

```
Best Training Accuracy: 0.8135457931768462
Best Testing Accuracy: 0.8124301987938352
Cross-val-score: 0.8131690261446602
```

I conclude from above that this model is not overfitting or underfitting since the training and testing scores are close. The testing accuracy of the model is 81.2%. The cross-val-score being close to the testing score indicates that the testing set is a valid representation of the data.

Since the logistic regression model does not provide predictive power as high as I would like, I will attempt a more complex Decision Tree model next.