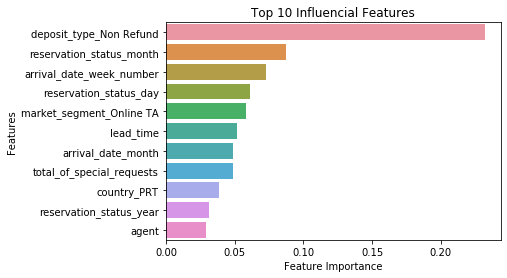
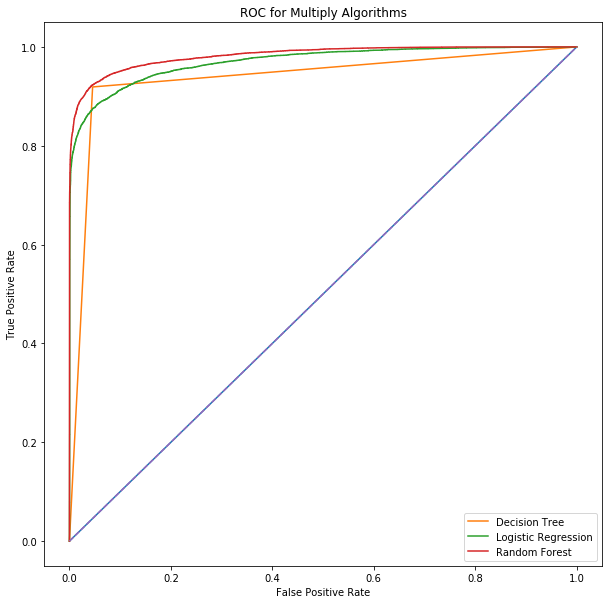
Machine Learning Report

In this part of the project, I am comparing three types of machine learning models to find the one that has the best performance in predicting or classifying a client cancel their hotel reservation.

To feed the data into models, I need to preprocess the data. First, I use the function get\_dummy from pandas to create categories for string inputs. The next step I have done is feature selection to reduce model dimensions and have a better result. I separated the data into a training set and test set for evaluation later and avoid overfitting. Then I feed the training set into DecisionTreeClassifier and look for important features using the attribute ‘feature\_importance’. The result I got back is that only ‘reservation\_status’ have 100% contribution and the rest have 0% contribution to the feature importance. This represents that the feature ‘reservatio\_status’ is leaking the information about whether a client is canceling their reservation. This makes sense and I must remove this feature to build my model. After removing the feature and rerun the codes, I got a list of feature importance. I dropped all the features that have 0 contributions to the model and keep the rest. Here are the top ten most influential features:

I feed the necessary features into DecisionTreeClassifier and calculate the AUC score; the result is about 0.93665. For logistic regression, the default setting gave an AUC score of 0.96977 and with tuning the C value using grid search, the AUC score improved a little at a value of 0.97776. Lastly, for RandomForestClassifier, the default setting returns an AUC score of 0.97534. After tuning the parameter using a randomized grid search, the AUC score is 0.983285. As a result, the RandomForestClassifier has the best performance in predicting the probability of a client canceling their reservation. Here is a ROC curve plot of the three algorithms:

In conclusion, to make the model better I think I should add holiday labels to each day because we see some influences from reservation status date. We can also maybe look at the agent-client pool and compare it to people with a different agent or no agent to better understand why that is influencing the cancelation of a reservation. For recommendation, because non-refund type room contribute to almost one-fourth of the feature importance, the hotel manager should increase their non-refund type rooms to reduce reservation cancelation.