GREG: R OUTPUT AND COMMENTARY FOR FINAL PROJECT

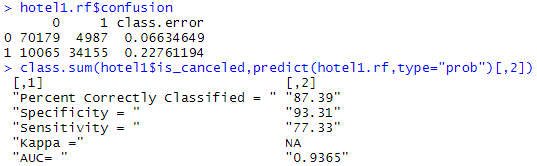
RANDOM FORESTS

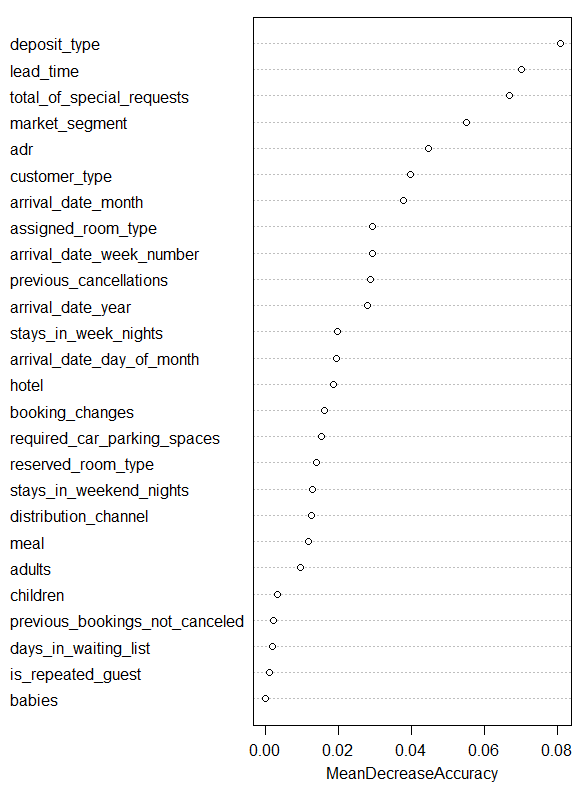
**We first applied random forests to the original dataset. We dropped the country, agent, company, and reservation\_status\_date columns because randomForest in R does not handle categorical variables with more than 53 levels. We also dropped reservation\_status because it was the same as the is\_canceled response variable. Just 4 observations with missing values were dropped as well. We then refit random forests with subsets of the important variables to see the changes in accuracy.**

**After analyzing the combined data, we applied random forests to the two hotel types separately to discover their differences. We also produced some two-way frequency tables to examine important variables for the combined and separated datasets. This was another good way to find differences between the resort hotel and the city hotel.**

**Finally, we applied random forests using hotel type as the response variable (as in the Nest data homework) to identify where misclassifications were occurring.**

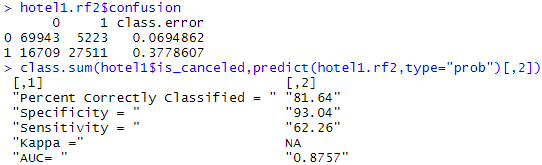
Combined data random forests result:





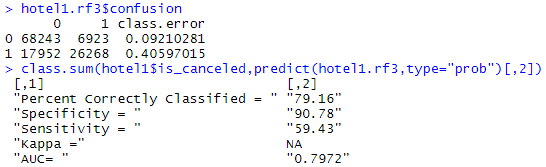
**The overall accuracy for random forests is very good, but we have less than ideal sensitivity. We chose to refit random forests using the top 7 variables, and then using the top 4 variables.**

7-variable model:



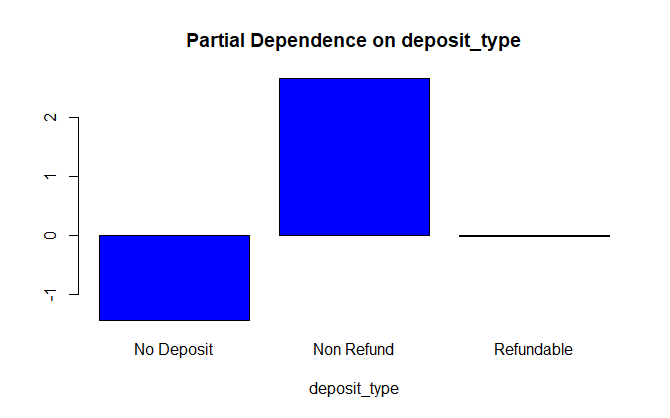
**The overall accuracy decreased by about 6 percentage points, and the sensitivity decreased drastically.**

4-variable model:

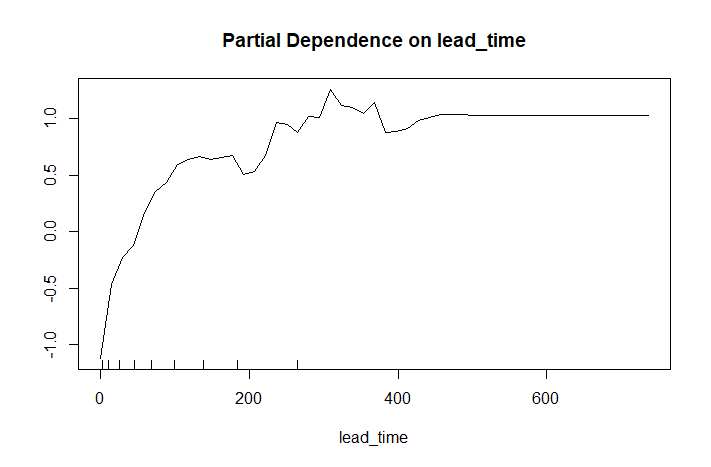


**Each metric decreased further by a few percentage points.**

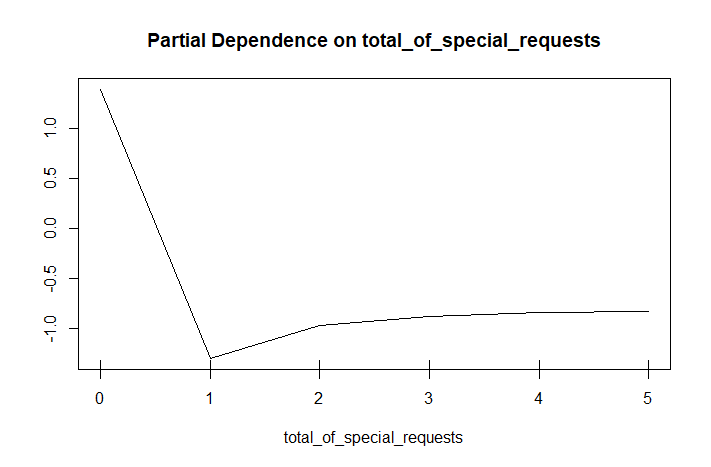
Partial Dependence Plots for Combined Data (Top 8 Variables, and Hotel):



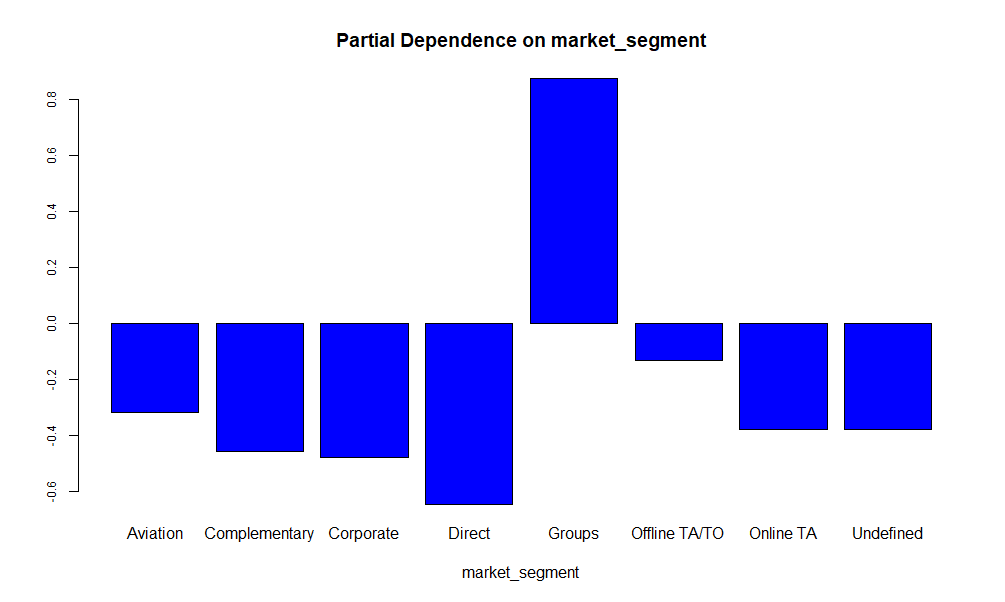
**Bookings were much more likely to be canceled if deposit\_type was Non Refund.**



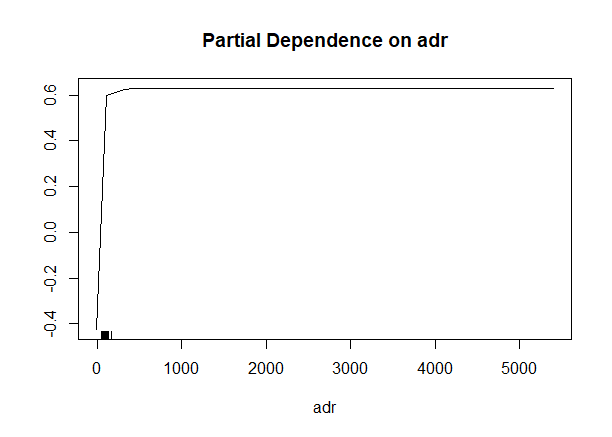
**Cancelation was more likely for higher lead times between reservation and arrival. The likelihood changes most between about 0 and 100 days.**



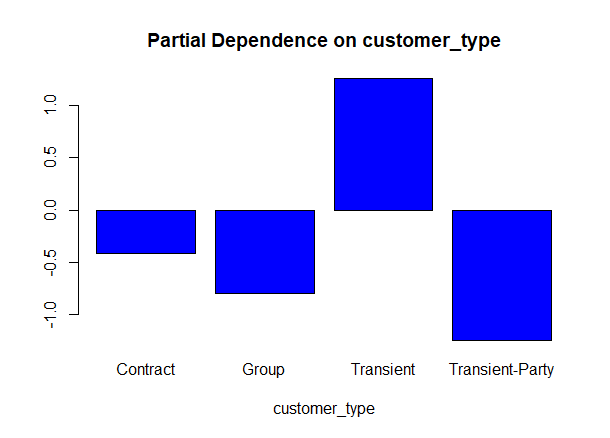
**Cancelation dropped drastically where at least one special request was made.**



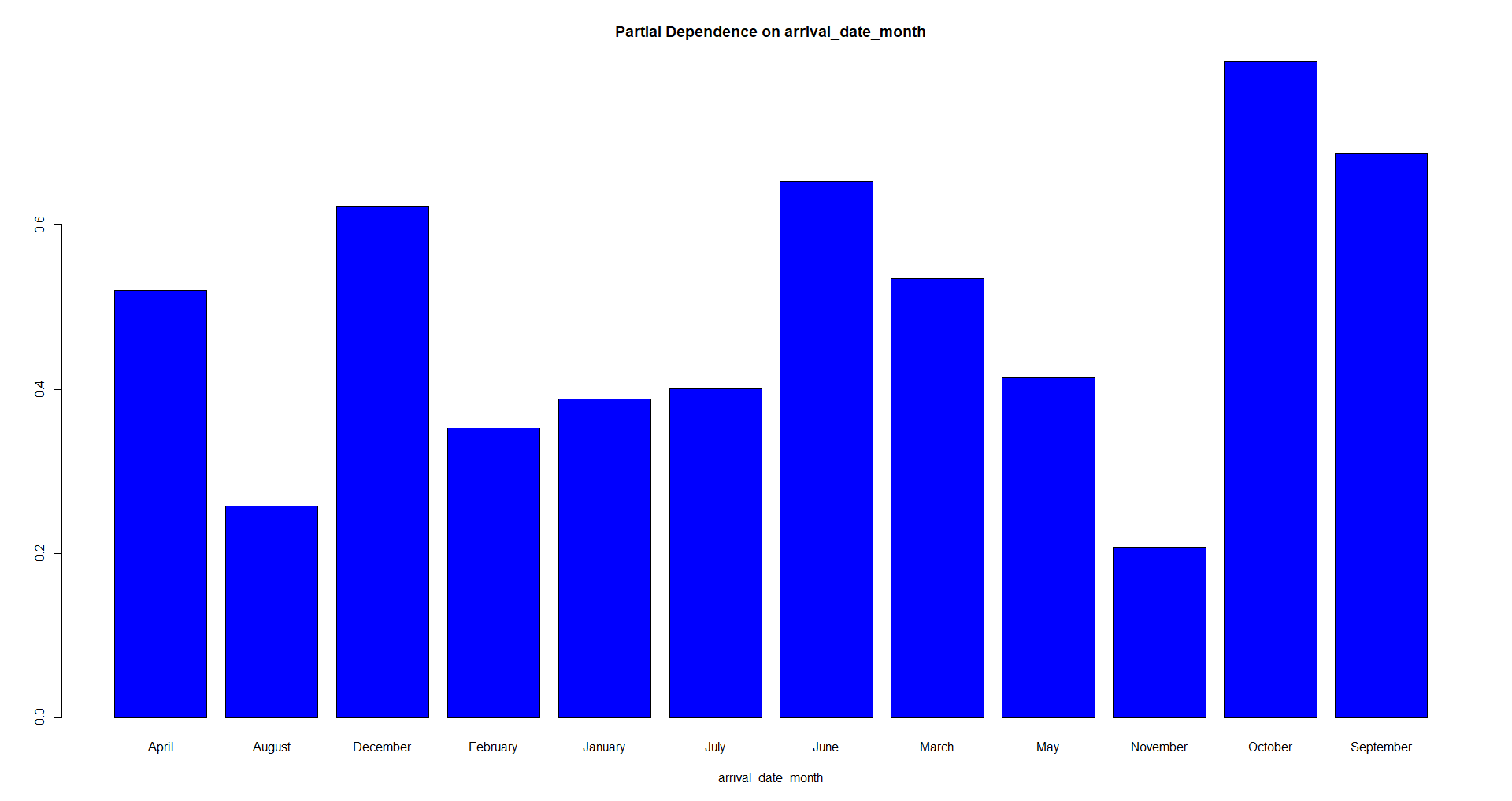
**Groups were much more likely to cancel than any other market segment.**



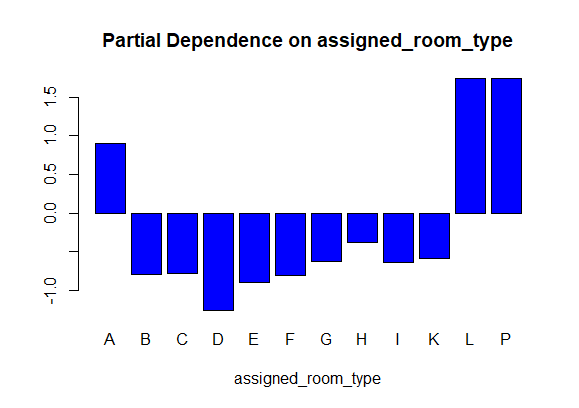
**Cancelation was much more likely beyond an average daily rate of about 100 to 200.**



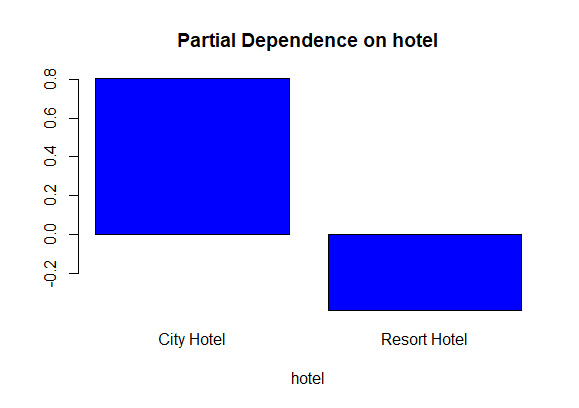
**Transient customers were more likely to cancel – this was also shown in classification trees.**



**For the combined data, cancelations were more likely in October.**



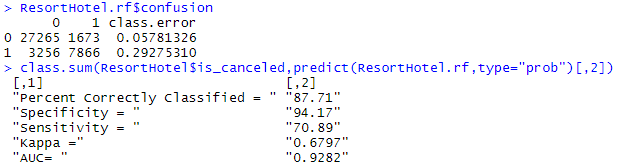
**Room type A was more likely to be cancelled, and there were very few observations with room types L and P, which were all cancelled.**

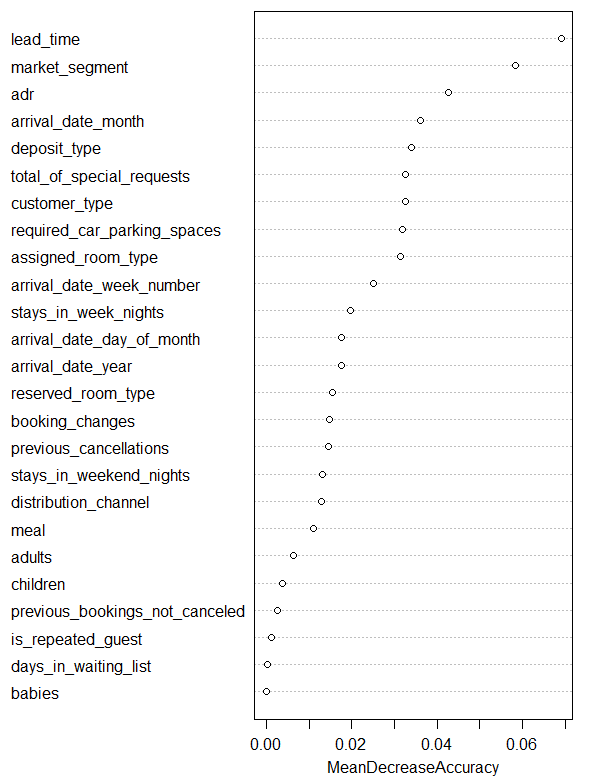


**The city hotel had a higher cancelation rate than the resort hotel.**

Separate Hotel Types Random Forests Results:

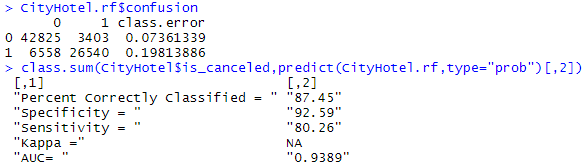
Resort hotel random forests result:

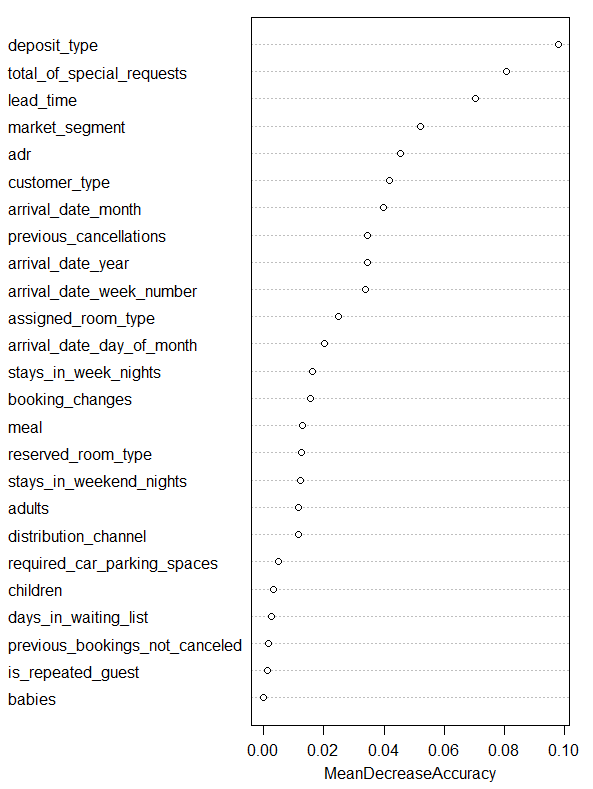




**Random forests on the resort hotel had the highest overall accuracy and specificity, but lower sensitivity. The important variables shifted around a lot, with lead\_time now at the top and deposit\_type fifth. This suggests the resort hotel is different from the city hotel.**

City hotel random forests result:

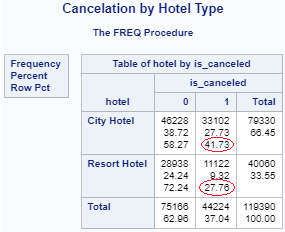




**We got slightly higher accuracy and sensitivity with the city hotel, and the most important variables stayed relatively the same as the combined data.**

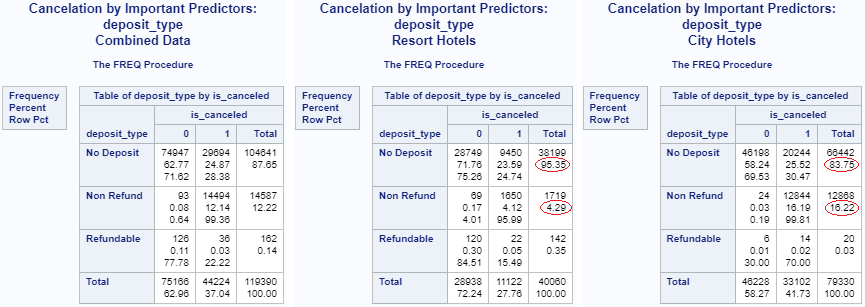
Frequency Tables for Important Variables

Cancelation by hotel type:



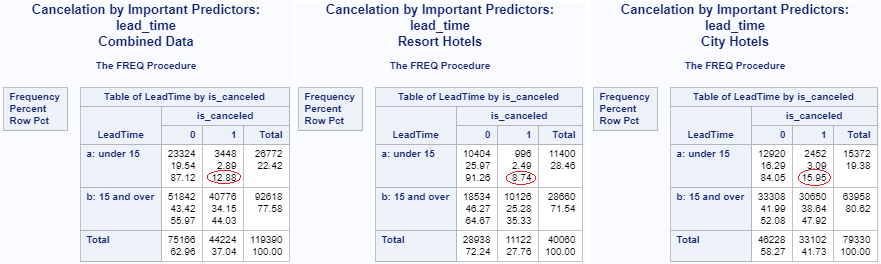
**The two hotels had very different overall cancelation rates (27.76% versus 41.73%), which is good evidence for treating them separately.**

deposit\_type:



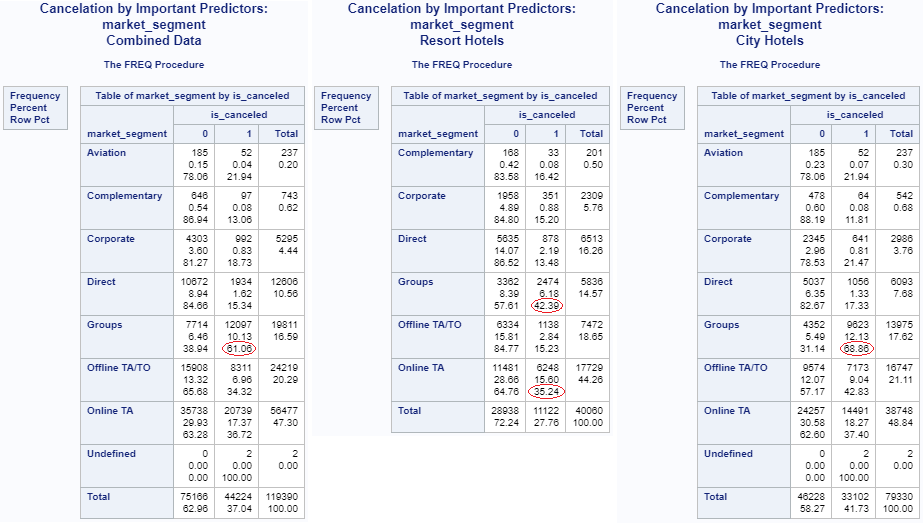
* **In all three cases, the Non Refund deposit type had extremely high cancelation rates. This was surprising because Non Refund indicates a deposit was made in the value of the total stay cost. One explanation may be that only 4.29% of resort hotel bookings were Non Refund, and only 16.22% of city hotel bookings were Non Refund, so it was a less typical option.**
* **This also highlights a difference between the resort hotel and the city hotel. As shown in the variable importance plots, deposit\_type is less important for the resort hotel because over 95% of resort bookings were No Deposit, and slightly less Non Refund bookings were canceled.**

lead\_time:



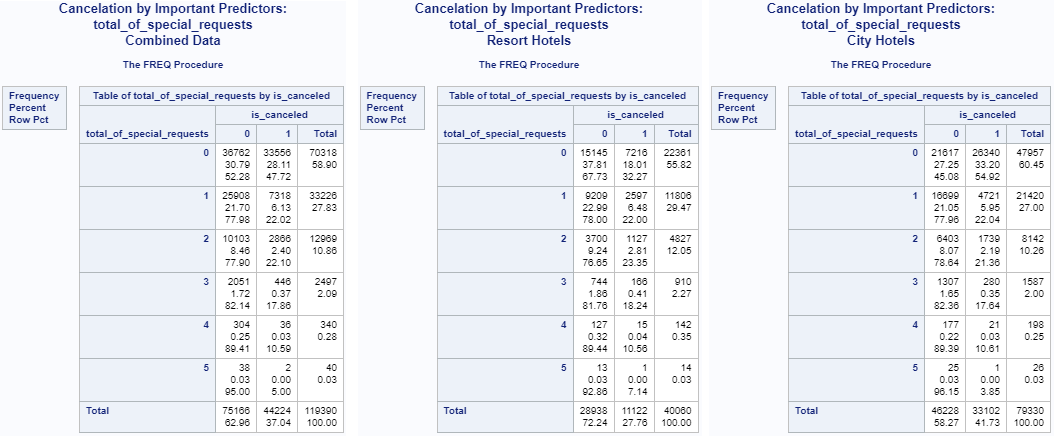
**In all three cases, cancelation rates were much lower when reservations were made less than 15 days before arrival. The grouping cutoff of 15 days is based on the lead\_time split from the Resort Hotel classification tree. The variable importance plot above also shows lead\_time as the resort hotel’s most important variable.**

market\_segment:



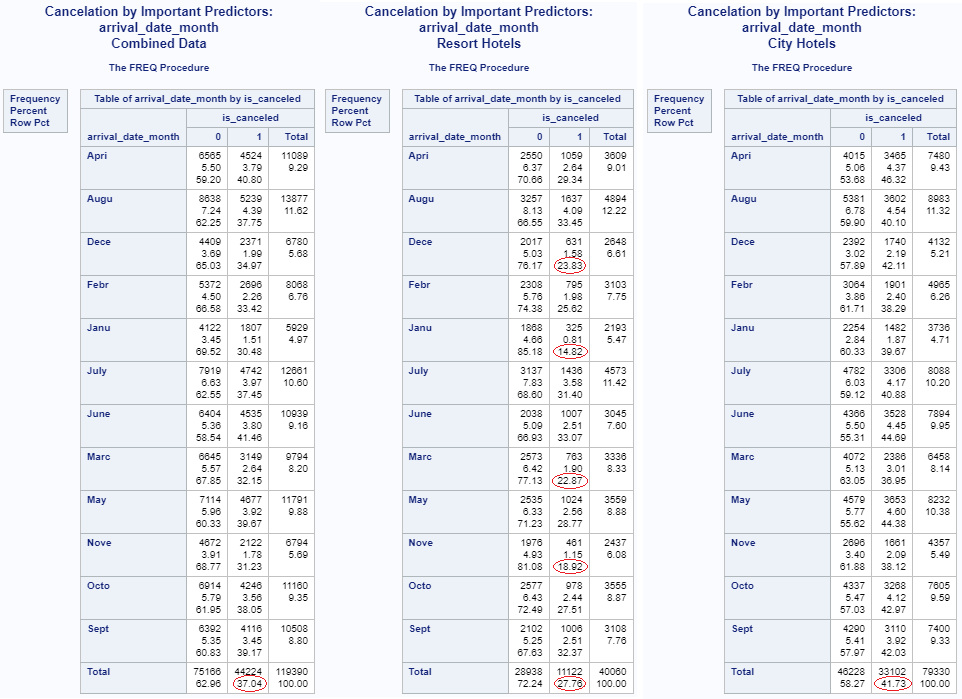
**In all three cases, Groups had the highest cancelation rates. For the resort hotel, the ‘Online TA’ segment had a higher than average cancelation rate as well.**

total\_of\_special\_requests:



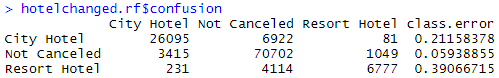
**Cancelation decreased as special requests increased for the combined data as well as for the separate hotel types.**

arrival\_date\_month:



**For the resort hotel, November, December, January, and March had significantly lower cancelation rates than the average of 27.76%. The combined data and the city hotels had less variation in monthly cancelation rates, so it is harder to predict the best times of year for them. This agrees with the variable importance plot for resort hotels, which shows arrival\_date\_month as more important.**

Out-of-Bag Confusion matrix for RF with HOTEL as Response:



* **Most of the misclassifications were the two hotels being confused for Not Canceled. The hotels were not confused for each other very often. This is further reason to treat the resort hotel and the city hotel differently.**
* **This confusion matrix also shows the lower sensitivity and higher specificity we had with our other random forests fits.**

Random Forests Conclusions:

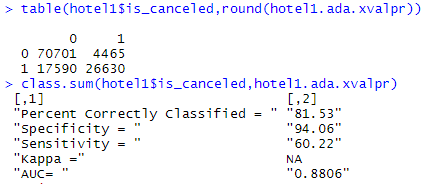
* **Random Forests gave us very good predictive accuracy for is\_canceled compared to the other methods. It identified deposit\_type, lead\_time, total\_of\_special\_requests, and market\_segment among the most important variables for prediction cancelation.**
* **Our reduced models had decreased accuracy, but the 7-variable model was not bad apart from the decreased sensitivitiy.**
* **The resort hotel is different from the city hotel because:**

1. **It had a much lower overall cancelation proportion.**
2. **It had different proportions of deposit types.**
3. **lead\_time was a better cancelation predictor for the resort hotel.**
4. **It had more variation in cancelation rates by month, which made arrival\_date\_month a better cancelation predictor.**

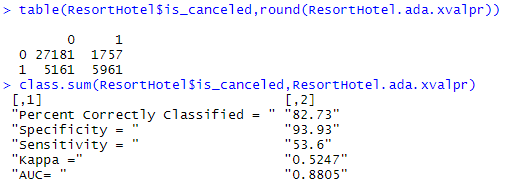
* **In each case examined, the Groups market segment had the highest cancelation rate.**
* **In each case examined, more special requests resulted in less cancelation.**

ADABOOST:

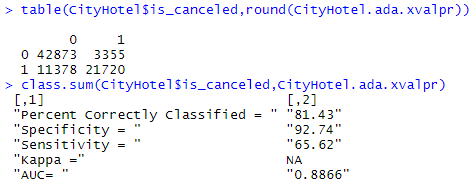
Combined Data:



Resort Hotel:



City Hotel:

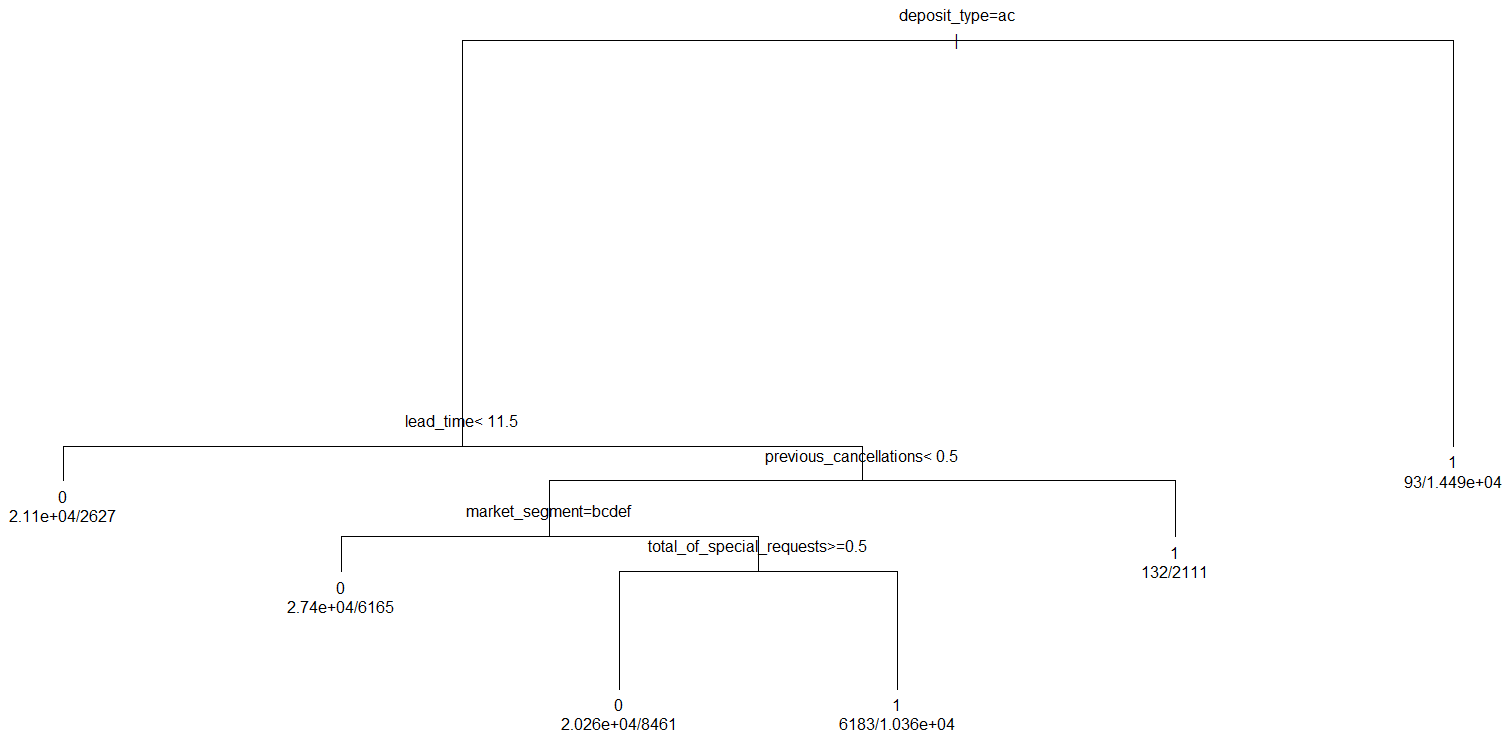


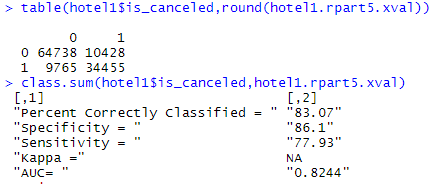
Adaboost Conclusion:

**Adaboost gave similar results for each of our three cases, and it was slightly less accurate than random forests.**

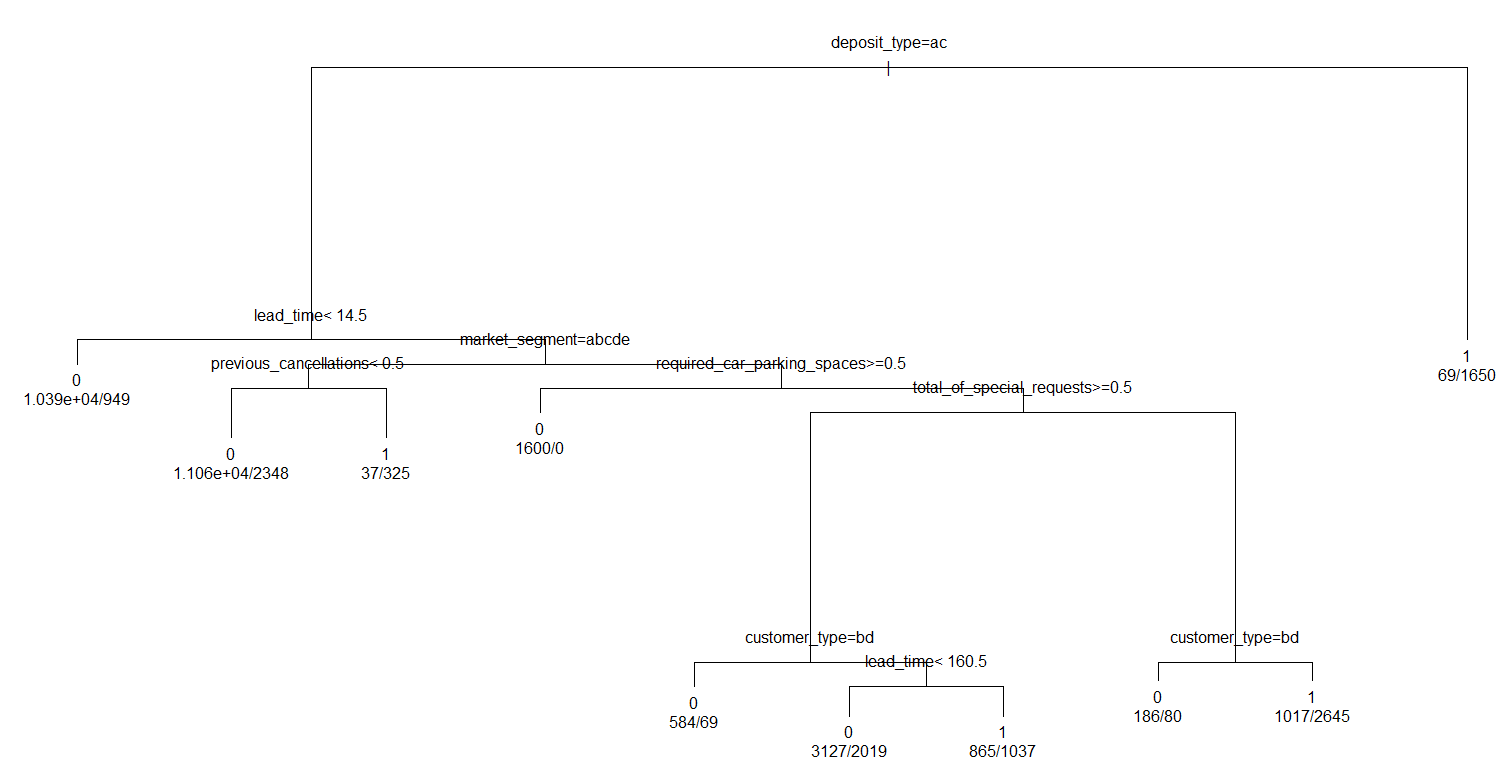
CLASSIFICATION TREES

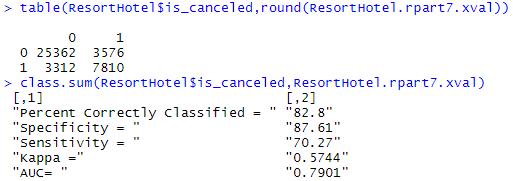
Combined Data, 5 splits



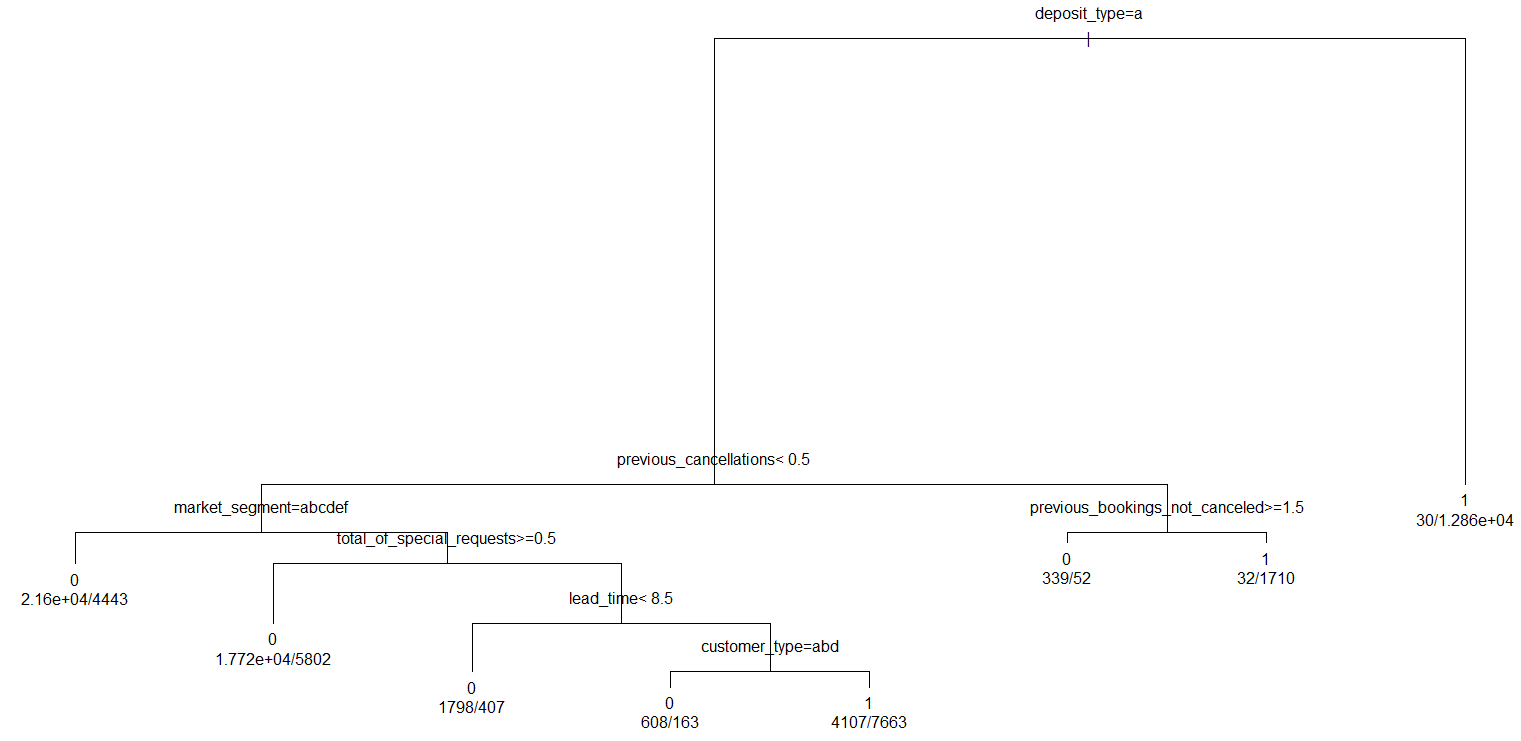


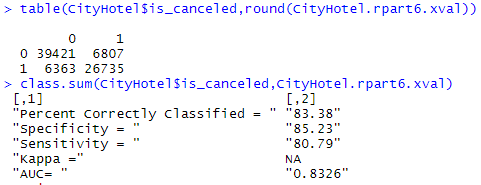
Resort Hotel Only, 7 splits





City Hotel Only, 6 Splits





CROSSVALIDATED CONFUSION MATRIX FOR HOTEL AS RESPONSE