# Retention Modeling at Scholastic Travel Company(A)

**Prediction Task:**

**To predict which customers would book with Scholastic Travel Company in the 2013–14 school year (fall 2013 to spring 2014).**

**Methods Used:**

**To build a model that took the data available as of spring 2013 to make this prediction. To build such a model, however, we would need to replicate the 2013–14 prediction task on the available data. This meant that for training his model, we would use the data from the 2012–13 school year—which showed whether a certain group had been retained or not—and try to predict based on the client-profile information as of the end of the 2011–12 school year.**

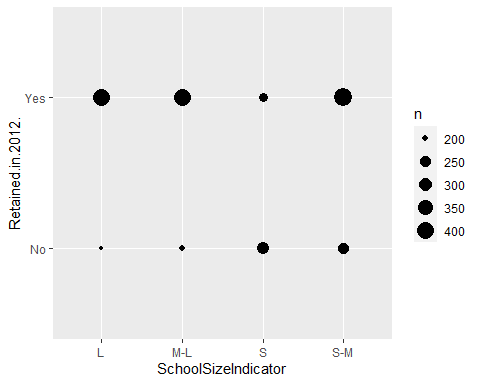
**Importing and cleaning the dataset (Code and Output Page 1-4)**

**Steps followed:**

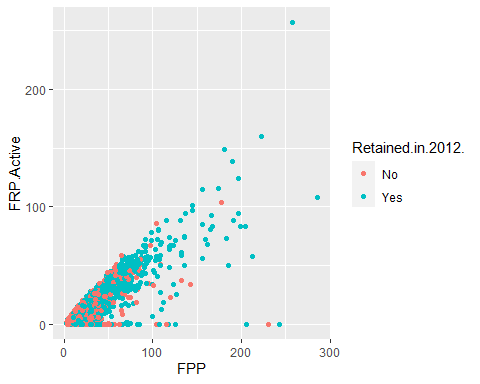
* Removing the unnecessary columns by looking at the initial stage
* Matching two columns as they have similar data and removing one of the columns
* Converting character into numeric variables
* Imputing NA values with mode of the column for categorical variables
* Imputing NA values with calculated means for numerical variables
* Mutating various numeric columns to factor
* Mutating all the character variables to factors

**Exploratory Data Analysis (Code and Output Page 4-13)**

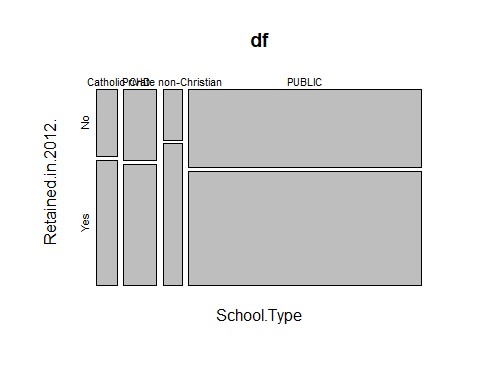
Insights:



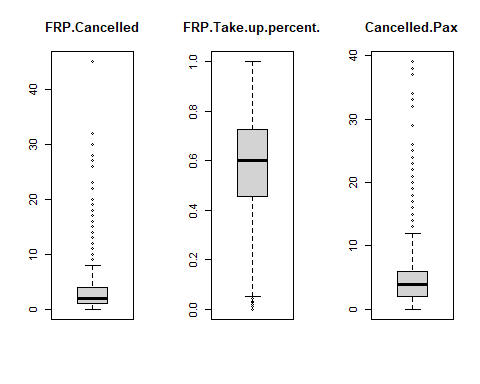
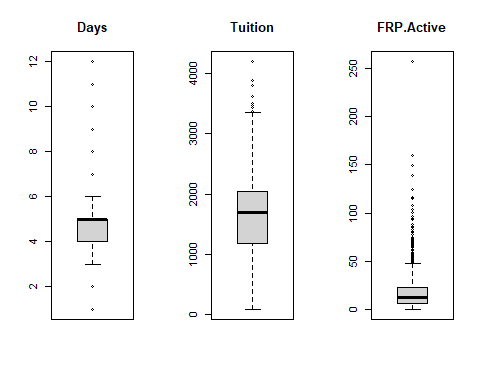
* We can infer from graph 1 that S-M size schools have the highest retention rate.

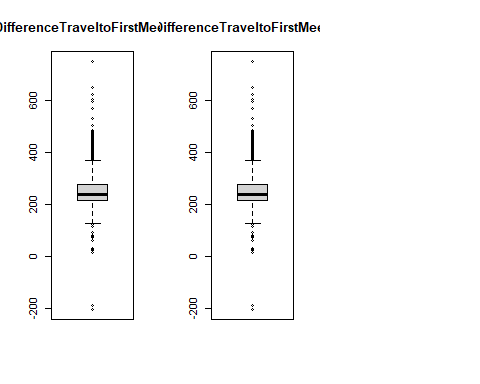
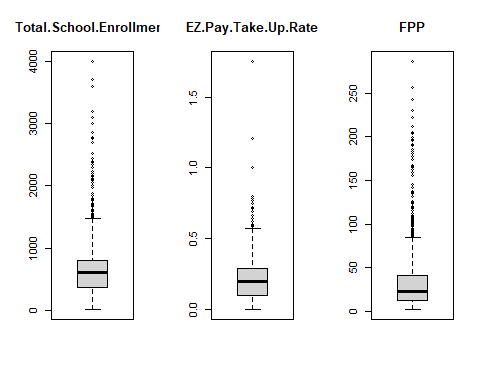


* The scatterplot shows all the data points and differentiates between being retained or not retained. We can observe that as the number of full paying particpants increases the number of people who have taken the insurance and have been retained for the next year.

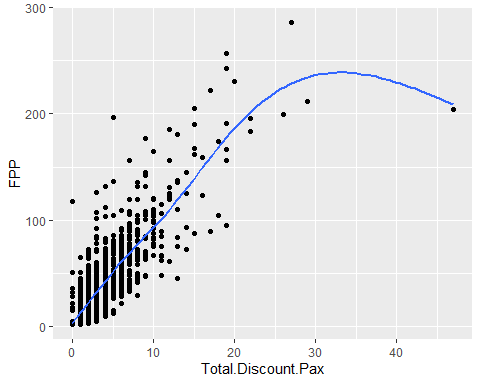


* The public school type has retained STC the highest and Private non christian schools retained the STC least when compared to all the other school types.

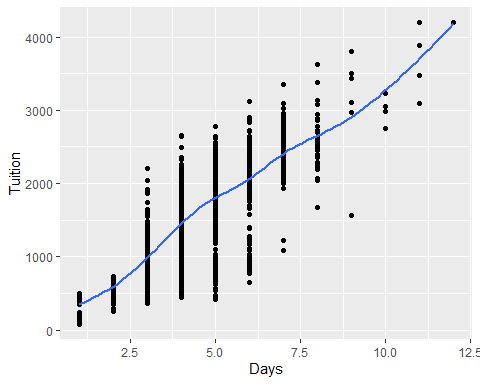




* We have plotted the boxplots of all numerical variables to view where the median lies for each variable.



* We can infer from curvilinear relationship that the discount paid by fully paying participants steadily increased and then started decreasing.



* As the number of days the group on program increases, the price that costs for program also goes up. To verify the above, we have found out the correlation between days and tuition. As we can see there is a high correlation between these two variables.

**Finding the important variables using random forest and eliminating the variables that have a mean gini decrease of less than 10. (Code and Output Page 13-15)**

**After removing the following variables, we have a data frame that finally consists of 35 important variables according to our analysis**

**Performance Metric Used- Recall**

**Reason:** If STC is actually being retained by customer, but they are predicted as not retained, it will be expensive. Hence, we consider recall to be a performance metric for this dataset as it will be crucial to reduce the number of False negatives (actually retained but predicted as not retained).

**Building different Machine Learning Models:**

**Decision Trees (Code and Output Page 16-30):**

Construction of various decision trees by experimenting with different train and test (50:50,70:30,80:20) splits.

**Result:**

According to the decision trees that have been constructed above, we can conclude that 70:30 split with the pruning parameters gives us the best recall.

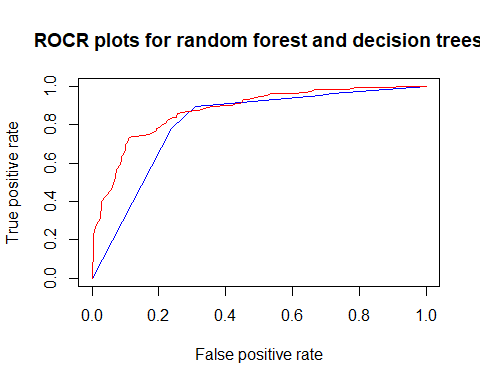
**Random Forest Model (Code and Output Page 16-31)**

Constructing random forests with different ntree values and finding the best mtry for each model to derive a single model with the best performance

**Result:**

According to the randomforest models constructed above, the randomforest model1 is considered as it gives us the better recall.

**Plotting the ROC curve to see which model performs better (Page 32-34):**



The red and the blue lines indicate the ROC curve for Random Forest and Decision Tree respectively. From this graph we can state that random forest has a better performance on this dataset when compared to decision tree.



From the auc that we calculated above for both the models, we have inferred that RandomForest model has better auc when compared to decision tree model. By this, we can say that RandomForest model performs better than decision tree model.

**Cross Validation (Code and Output Page 34):**

Performing 10-fold cross validation on the dataset for both random forest and decision tree to see the effect of the performance and calculate the weighted averages.



When considering 10- fold cross validation, we derived that decision tree has a higher recall when compared to random Forest model.