Group 1 Case 4

Scholastic Travel Company is an educational tourism firm and customer retention is of key importance for it to grow sustainably. This is why the management has started a new data initiative where they want to develop a model that helps them predict whether or not a customer would book again in the future. The dataset provided has all the known information about the previous clients including multiple numerical and categorical variables such as total school enrollment, tuition, income level, travel and grade type etc.

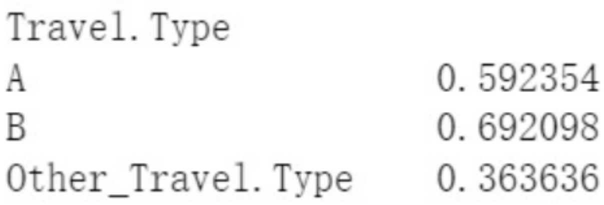
Since it's a huge dataset with multiple variables and records, prior to fitting any models, we have cleaned the data to improve our operability. As part of the cleaning exercise, we also reduce the level of variable outcome by aggregating. We started by aggregating the categorical variables 'From.Grade' and 'To.Grade' into 'Elementary', 'Middle', and 'High' class categories. Next, the 'Days' variable was aggregated into 'Short_Length', 'Middle_Length', and 'Long_Length' trip categories. Similarly, states were grouped into 5 categories: NorthEast, MidWest, South, West, and missing. The above mentioned variables were converted to 'Category' type along with all of the other categorical variables in the dataset. Missing values were filled with either 'Missing' or 0 for character or numeric variables, respectively. Some of our categorical variables had rare levels, i.e. categories that only appeared less than 20 times in our data (arbitrary cutoff chosen by us). These categories were rolled up into an 'Other' category.

Next, an exploratory data analysis was conducted to determine which variables could be disqualified as predictors. For example, we explored the relationship between the trip's length, whether the customers were existing or new etc against their retention. This helped us refine our predictors.

Part A:

In order to choose our predictor variables for our models, we first filtered out the variables for which information may not be available before the trip, such as the *FPP* and *Total Pax*, etc. For the first kNN model we settled on the following predictors: Tuition, Total.School.Enrollment, From.Grade, To.Grade, Group.State, Is.Non.Annual., Days, Travel.Type, Parent.Meeting.Flag, School.Sponsor, Income.Level, SingleGradeTripFlag, SPR.Product.Type, SPR.New.Existing. To find the best nearest neighbor model we ran 24 iterations with 'k' ranging from 1 to 24. A k value of 24 yielded the best accuracy score of .786. However, a model using 24 nearest neighbors would overfit the training data and minimize the local data structures in the dataset. Due to these considerations and a desire for a simpler model we settled on a model with 8 nearest neighbors. This model yielded an accuracy score of .7764 on the validation data.

In an attempt to further improve our accuracy score, we decided to fit another kNN model this time without 'Travel.Type'. We removed this variable because from the table given below, it was observed that the mode of transportation for the trip was not very relevant to predicting retention of STC. This proved to be a correct assumption as our accuracy on the validation data shot up to .7816.



```
Travel.Type
A                      0.592354
B                      0.692098
Other_Travel.Type      0.363636
```

For similar reasons we also ran a model without the ‘Group.State’ variable but the accuracy score for that model went down, and therefore, we took our second KNN model as the best KNN model.

Next, we moved on to classification trees. For our first classification tree we decided to keep the same variables as for our second kNN (‘Travel.Type’ was dropped). No limits on tree size and purity were placed – the tree was allowed to grow to its deepest length – 639 nodes. This large tree yielded an accuracy score of .7032 on the validation data, which was lower than what we got on either of the kNN models. To improve our tree’s predictive performance, we used a grid search and set tree pruning parameters accordingly (max depth 5, minimum samples split 50, and the minimum impurity decrease 0.001). This process helped us prune our tree and reduce it down to a simpler and more efficient tree with 39 nodes and an accuracy score of .79 on the validation data.

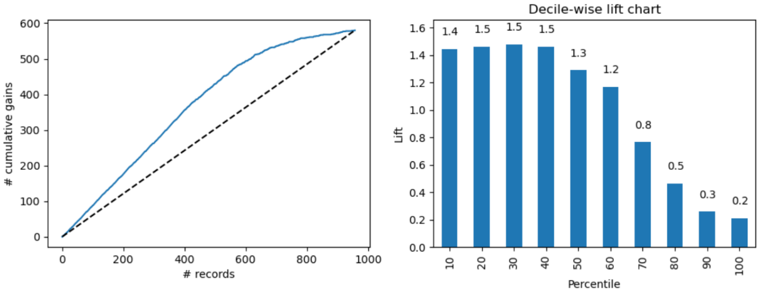
We experimented with our predictor variables to improve our classification tree as well, but our best tree was the one we pruned keeping the same predictor variables as our second kNN model. This was validated by very high boosted tree (0.799) and random forest (0.78) accuracy scores as well.

Part B:

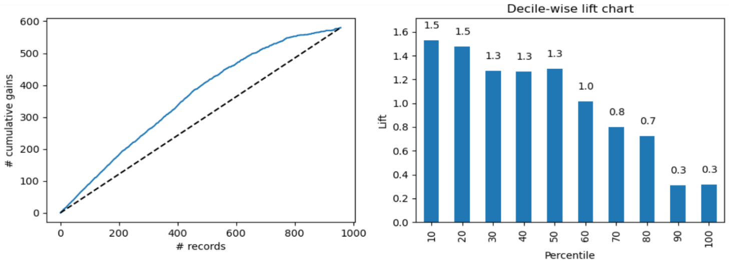
We ran two logit regressions model, one with predictors aligned with our previous best classification tree model and one with all the predictors from the dataset (excluding the variables for which information could only be obtained after the trip). The predictive accuracy on validation data of the first logit regression (0.794) was higher than the second one (0.742).

Another way of comparing the predictive performance of logit models is to compare cumulative gains and lift charts. A general rule of thumb for cumulative gains chart is that the further away the cumulative gains curve from the diagonal base line, the better the model is doing in separating records with high value outcomes from those with low value outcomes. Just as the accuracy score for the two models is close, the outcomes of their charts are also similar. However, it can be observed that the gains chart for the first model is more lifted than the second one, suggesting, for the top 30% of the records suggested by our logit model, we will have 1.5times more customer retention than with randomly chosen 30% (700) records.

1st Logit Regression Model:



2nd Logit Regression Model:



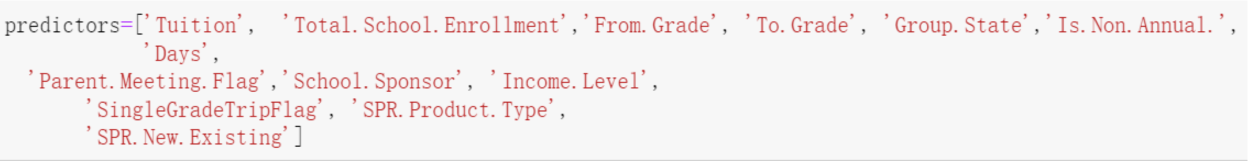
Part C:

Model Type	**Predictive Accuracy on Validation Data**
k-NN	0.789
Classification Tree	0.799
Logit Regression	0.794

The three models have comparable accuracy scores with classification trees having the highest. A good model is defined by its performance on validation dataset in the domain of data mining. Accuracy is one of the most important indicators, but robustness and operability are also important especially when accuracy scores are so commensurate. kNN is a non-parametric model and it's also a lazy learner, meaning it runs over and over again and can lead to time inefficiency and repetition. On the other hand, whilst classification trees are relatively easy algorithms, they are unstable. A different partition could lead to a different tree and prediction altogether. That problem is taken care of with the help of the more robust boosted tree algorithm, but again that leaves you with little information on the variables.

Logistic regression is easy to interpret and implement especially on large datasets with many records. It not only provides a measure of how important a predictor is but also if the association is negative or positive. Hence it can be a good prediction model in this dataset, especially when the dependent variable is discrete.

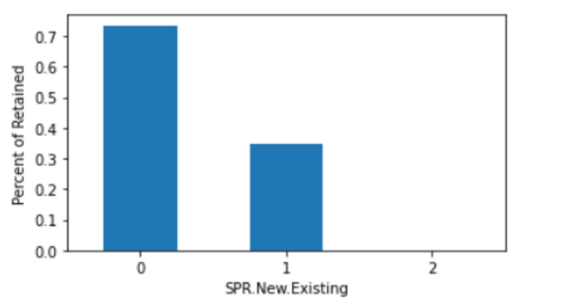
Part D:

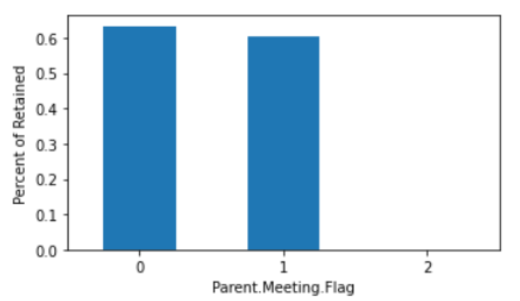


```
predictors=['Tuition',  'Total.School.Enrollment','From.Grade', 'To.Grade', 'Group.State','Is.Non.Annual.',
            'Days',
  'Parent.Meeting.Flag','School.Sponsor', 'Income.Level',
       'SingleGradeTripFlag', 'SPR.Product.Type',
       'SPR.New.Existing']
```

We chose the above predictors for our final model based on predictive accuracy and analyses above. We chose our variables based firstly on whether the information regarding that predictor would be available before the trip, resultantly variables like ‘Cancelled.Pax’ were dropped in the first place.

Next we also did some preliminary relationship exploration to see which variables were relevant. So as it can be seen from the images below, the variable defining existing vs. new customers turned out to be a good predictor, whereas parent meetings did not seem to have a lot of impact on retention.





We also considered the relevance of predictors based on random forest results, for example, Total school Enrollment and Tuition were some variables that were of high importance and were to be retained in the model. Thirdly, we played around with adding and removing several variables, and then compared the accuracy scores on the validation dataset for different models. In the end, we chose the more parsimonious model with a smaller number of relevant variables given the accuracy scores were similar.

Part E:

‘Is.Non.Annual._1.0’ is a categorical dummy variable indicating that the group from this school typically skips a year in between programs. Its logit coefficient is ‘-2.094’, which means that if a school typically skips a year in between programs, that school is less likely to participate in the STC programs in the 2013-2014 school year. In other words, having a school go from annual to non annual, holding all other variables constant, will decrease the log of odds by 2.094. This makes sense since the data points are from schools who participated in STC programs in the 2012-2013 school year; if they plan to join every other year, they won't participate in STC programs in 2013-2014.

'SPR.New.Existing_NEW' is a categorical dummy variable which means that the school has never traveled before with STC with few exceptions. Its logit coefficient is '-1.623', which implies that a school which has not cooperated with STC before is associated with a lower possibility of participating in the STC programs in the 2013-2014 school year. Hence, having a school go from an existing customer to a new customer who has never worked with STC before will reduce the log of odds by 1.623. This is a reasonable explanation because customers tend to be more careful when choosing a new product and for a company acquiring a new customer is anywhere from 5 to 25 times more expensive than retaining an existing one.

'Total.School.Enrollment' is a continuous predictor showing the total enrollment of the school. So bigger schools have higher numbers of enrollment whereas smaller schools have lower values on this predictor. This variable has a positive logic coefficient of '0.00045' which means a bigger school is associated with higher possibility of participating in the STC programs in the 2013-2014 school year. In other words, increasing the enrollment of a school by one unit will increase the log of odds for a school returning to STC the next year by .00045. This might be explained by the fact that the larger the school, the larger number of students might be interested in taking such a trip.

Part F:

Based on our final logit model, our prediction is as follows: we would like to use our logit model to predict whether a particular student group from Northeastern middle school with a total enrollment of 2000 students would repurchase STC services in the 2013 - 2014 school year.

Let us call this prospective school Hogwarts Intermediate Preparatory School, or HIPS for short. A student group from HIPS would like to travel annually and send one grade at a time for a trip of medium duration (5 to 8 days – 'Days' was converted into a categorical variable). There was already a parent meeting held, where the parents were briefed on a potential overnight trip to an east coast state. As this trip is deemed important for the education of the students, HIPS is willing to sponsor this trip financially. STC has done a preliminary analysis of the student body and concluded that HIPS is mostly composed of students from middle income families. Knowing all of this information we use our logit model to predict the probability that this school will book with STC in the 2013-14 school year. Our model predicts the probability that HIPS would book with STC at 97%, while the probability that they would not book at 3%. To compute the odds we divide the probability of the event occurring by the probability that the event does not occur (odds = p / (1-p)), this gives us odds of 32.3 for the school booking with STC. Based on such odds, HIPS has a huge potential to be a future retention customer.