BAN 620 CASE ASSIGNMENT 3

SHRIYA ARORA – sz9461

PALLAVI NAIR- fb4097

PRIYADARSHINI MADHUSUDANAN- hy1162

ADITYA MANE- yy5910

SAI KUMAR- lp2345

HARSHITHA ANANTHULA- pi3407

CAL STATE EAST BAY

STC CASE

Introduction: STC is a company that specializes in arranging excursions for students that are both educational and cultural. The company has been in operation for a number of years and evolved into a prosperous multi-million-dollar business. STC is renowned for its capacity to handle the numerous intricacies involved in group travel, including securing required paperwork, reserving lodging and transportation, and ensuring the safety and security of its clients.

Most of the excursions that STC organizes are compensated by parents, with professors or university officials in charge of planning the schedule and activities. STC keeps track of pre-trip conferences between educators and parents/students because they frequently reveal crucial details about the journey ahead. Additionally, the business gathers and keeps track of information regarding the travel party and the instructor or administrator who is in charge of planning it also collects feedback after the trip. Powell has easy access to a cloud-based database that houses all of this data.

The success of STC may be largely given to its expertise in managing the complications of group travel, as well as to its attention to detail and dedication to client satisfaction.

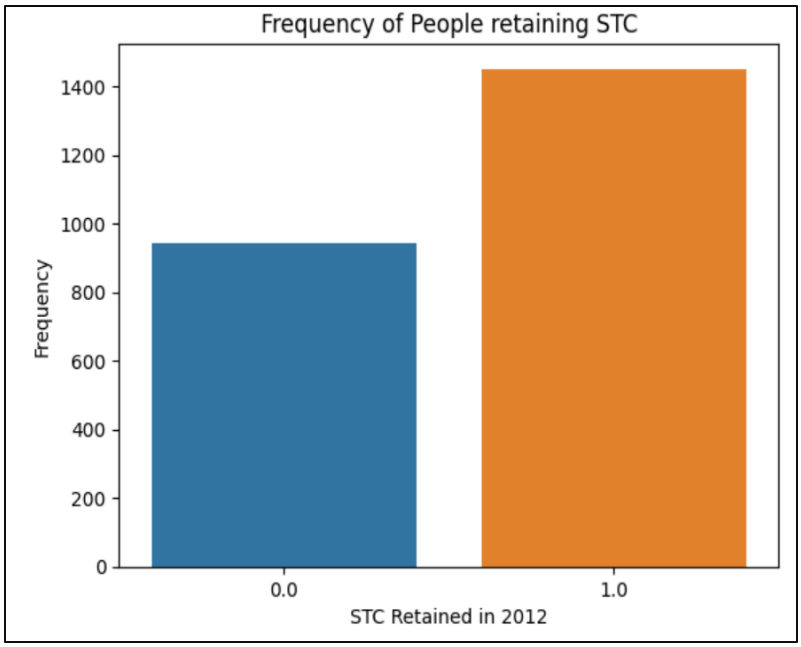
Goal: To develop various classification models using kNN, Trees algorithms and logit models to predict which customers are more likely to book a trip with STC.

a) Step 1: Cleaning the data:
Primarily, we converted the data type of categorical variables from object to categories.

In addition to that, we created dummy variables for the categorical variables and grouped the uncommon categories into one dummy variable.

To handle missing values in numerical variables, we replaced them with the median value. Whereas, in categorical variables, we have imputed the most common categories in the place of missing values

Data Source and Initial Analysis: The data was a part of HBR coursepack. The file contains data on 2392 customers. We have built a model based on customers retained by STC for the year 2012. The data consists of 55 variables. Some of the variables are School enrollment, which grade and the nature of holiday(annual/non-annual). Among these 2400 customers, around 1400 (= 58.3%) have been retained in the year 2012.



Step 2: Partition the data into training data and validation data and set the seed value.

Data set: Primarily, we divided the data set into training set: 60% and validation set: 40%

Step 3: Introduction of New customer for the Knn Model with Retained in 2012 as the dependent variable

	FRP.Cancelled	FRP.Active	Cancelled.Pax	FPP.to.PAX	From.Grade_Imputed	Is.Non.Annual._Imputed	Days_Imputed
0	2	12	4	0.9	4.0	0	3.0

School.Sponsor_Imputed	SingleGradeTripFlag_Imputed	SchoolSizeIndicator_Imputed	DepartureMonth_Imputed
1.0	1.0	3	4

SPR.New.Existing_Imputed	SPR.Product.Type_Imputed	Total.School.Enrollment	SPR.Group.Revenue	FPP.to.School.enrollment
1	6	2	2000	1

Step 4: Transform the dataset by initializing normalized training, validation and complete data frames and scale the data.

Step 5: Knn Model for Retained in 2012 with k=9 for a new customer.

Predictors: We have used a model consisting of 9 variables (FRP.Active, Total.School.Enrollment, SPR.Group.Revenue, FPP.to.School.enrollment, FPP.to.PAX, From.Grade, Is.Non.Annual., SingleGradeTripFlag, SPR.New.Existing) as predictors and Retained in 2012 as the dependent variable.

Step 6: Finding the best k

A loop is set up that iterates over different values of K (in this case, from 1 to 14). The output is a table that shows the accuracy of the KNN classifier for different values of K. By examining this table, it is possible to identify the value of K that gives the highest accuracy **(K=9 with 79.51% accuracy)** on the validation data.

Predicting the new data for k=9

	k	accuracy
0	1	0.719958
1	2	0.694880
2	3	0.755486
3	4	0.741902
4	5	0.786834
5	6	0.775340
6	7	0.791014
7	8	0.779519
8	9	0.795193
9	10	0.789969
10	11	0.793103
11	12	0.788924
12	13	0.794148
13	14	0.792059

```
[0.]
Distances [[7.93923038 7.94573995 8.0866536  8.10164161 8.47906817 9.01660178
  9.02672134 9.06380129 9.10133338]]
Indices [[1379  214 1245 1244 1060 1157 1395  144  112]]
      zFRP.Active  zTotal.School.Enrollment  zSPR.Group.Revenue  \
648      0.423464                -1.368027           -1.203961
2256     0.134290                -1.352911           -1.362325
714     -0.386224                -1.539338            0.727168
706      0.886143                -1.327719           -0.525913
958     -0.212719                -1.504068            0.150597
848      1.985005                -1.171523           -1.045596
1360     0.886143                -1.262217           -0.073881
833      2.563353                -0.723092           -1.855870
1842     0.596968                -1.342834            0.110621

      zFPP.to.School.enrollment  zFPP.to.PAX  zFrom.Grade_Imputed  \
648                    5.104977     0.512330            -0.945120
2256                   5.340820     0.634821            -0.220629
714                    3.978170     0.656869            -0.220629
706                    4.829827     0.634821            -0.945120
958                    3.869158     0.982081             0.503862
848                    3.184539     0.156543            -1.669611
1360                   2.904850    -0.033377            -0.945120
833                    3.705640     1.383813            -1.669611
1842                   3.189268     0.494263             0.503862

      zIs.Non.Annual._Imputed  zSingleGradeTripFlag_Imputed  \
648                  2.362746                     -1.108269
2256                 2.362746                     -1.108269
714                 -0.423236                     -1.108269
706                  2.362746                     -1.108269
958                 -0.423236                      0.902308
848                 -0.423236                     -1.108269
1360                -0.423236                     -1.108269
833                 -0.423236                      0.902308
1842                -0.423236                      0.902308

      zSPR.New.Existing_Imputed  Retained.in.2012.
648                   -0.705629                0.0
2256                  -0.705629                0.0
714                    1.417175                0.0
706                   -0.705629                0.0
958                   -0.705629                1.0
848                    1.417175                0.0
1360                   1.417175                1.0
833                   -0.705629                1.0
1842                  -0.705629                1.0
```

For k=9, when we fit in the model and predict for the new customer data, we obtained the prediction provided above.

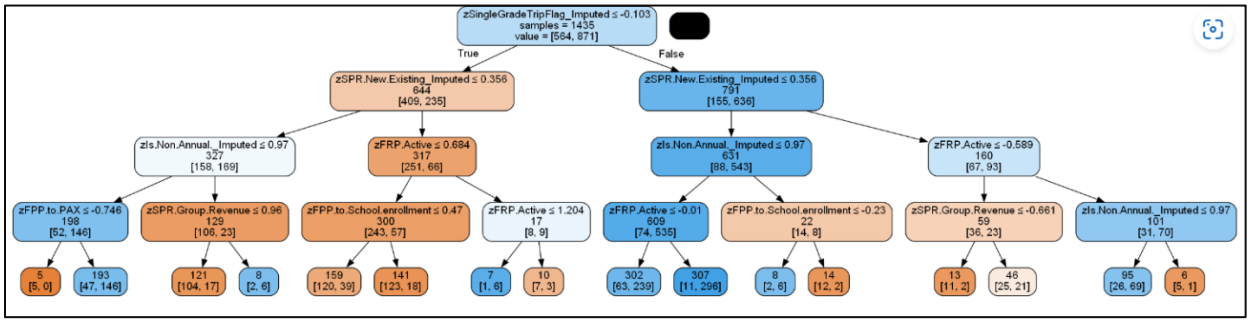
- This method returns two arrays: the distances between the new customer's data and the three closest points in the training data (stored in the "distances" variable), and the indices of those three points in the training data (stored in the "indices" variable).
- Finally, it shows the predicted target variable for the new customer and shows the rows of the training data that correspond to the nearest neighbors.

Classification Trees

In the decision tree model, we have used the **Grid search technique** such that we will be able to work systematically through all the hyper parameter values, to find the best result which helps in building the best model.

```
Initial score:  0.794425087108014
Initial parameters:  {'max_depth': 40, 'min_impurity_decrease': 0.001, 'min_samples_split': 100}
Improved score:  0.8034843205574914
Improved parameters:  {'max_depth': 4, 'min_impurity_decrease': 0.0009, 'min_samples_split': 10}
```

We have plotted the classification tree below using the grid search technique. We have obtained **78.89%** accuracy in the **validation data.** The Odds ratio obtained for the prediction accuracy of the validation data is 12.8.



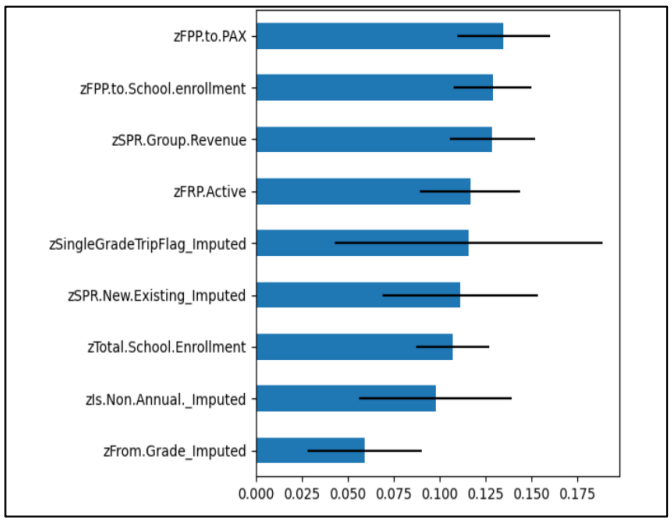

```
Confusion Matrix (Accuracy 0.8223)

       Prediction
Actual   0   1
     0 412 152
     1 103 768
Confusion Matrix (Accuracy 0.7889)

       Prediction
Actual   0   1
     0 256 118
     1  84 499
```

	Estimate	SE	LCB	UCB	p-value
Odds ratio	12.888		9.379	17.709	0.000
Log odds ratio	2.556	0.162	2.238	2.874	0.000
Risk ratio	4.751		3.853	5.858	0.000
Log risk ratio	1.558	0.107	1.349	1.768	0.000

Then we used the **Random Forest Classification model.** This method generally makes subsets from the dataset and builds decision trees and collates the output of the best models to obtain a decision tree.

We obtain the importance of each predictor variable employed in this model. Please find below the value and graphical representation of the importance obtained.

	feature	importance	std
5	zFrom.Grade_Imputed	0.059246	0.031059
6	zIs.Non.Annual._Imputed	0.097837	0.041570
1	zTotal.School.Enrollment	0.107227	0.019974
8	zSPR.New.Existing_Imputed	0.111075	0.042315
7	zSingleGradeTripFlag_Imputed	0.115630	0.072708
0	zFRP.Active	0.116674	0.027122
2	zSPR.Group.Revenue	0.128700	0.023030
3	zFPP.to.School.enrollment	0.128869	0.021035
4	zFPP.to.PAX	0.134742	0.025168



For this model, we have achieved an accuracy of **79.41%.** The Odds ratio obtained for the prediction accuracy of the validation data is 14.08.

Confusion Matrix (Accuracy 0.7941)

Actual \ Prediction	0	1
0	250	124
1	73	510

	Estimate	SE	LCB	UCB	p-value
Odds ratio	14.085		10.163	19.521	0.000
Log odds ratio	2.645	0.167	2.319	2.971	0.000
Risk ratio	5.338		4.258	6.693	0.000
Log risk ratio	1.675	0.115	1.449	1.901	0.000

In addition to that, we have used the **Gradient Booster Classification model,** this model generally builds sequential subset of models. This will always try to reduce the errors compared to the previous model. The accuracy we achieved with this model is **79.62%**

Confusion Matrix (Accuracy 0.7962)

Actual \ Prediction	0	1
0	259	115
1	80	503

	Estimate	SE	LCB	UCB	p-value
Odds ratio	14.161		10.259	19.546	0.000
Log odds ratio	2.650	0.164	2.328	2.973	0.000
Risk ratio	5.047		4.073	6.254	0.000
Log risk ratio	1.619	0.109	1.404	1.833	0.000

b) Logit model:

Step 1- As mentioned before, we perform data pre-processing and splitting of training and validation data.

Step 2- Build the model: We have used 6 variables as predictor variables and Retained in 2012 is considered as the outcome variable.

The variables used for prediction are: FRP.Active, FPP.to.School.enrollment, FPP.to.PAX, Is.Non.Annual., SingleGradeTripFlag, SPR.New.Existing.

Please find below for the intercepts and coefficients obtained.

Generalized Linear Model Regression Results

Dep. Variable:	Retained.in.2012.	**No. Observations:**	2392
Model:	GLM	**Df Residuals:**	2386
Model Family:	Binomial	**Df Model:**	5
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1104.4
Date:	Wed, 26 Apr 2023	**Deviance:**	2208.7
Time:	09:51:23	**Pearson chi2:**	2.55e+03
No. Iterations:	5	**Pseudo R-squ. (CS):**	0.3403
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
FRP.Active	0.0389	0.005	7.725	0.000	0.029	0.049
FPP.to.School.enrollment	-0.5622	0.668	-0.842	0.400	-1.870	0.746
FPP.to.PAX	0.2949	0.134	2.198	0.028	0.032	0.558
Is.Non.Annual._Imputed	-2.4945	0.174	-14.374	0.000	-2.835	-2.154
SingleGradeTripFlag_Imputed	1.1184	0.111	10.082	0.000	0.901	1.336
SPR.New.Existing_Imputed	-1.5169	0.113	-13.420	0.000	-1.738	-1.295

```
intercept  -1.824349468222632
       FRP.Active  FPP.to.School.enrollment  FPP.to.PAX  \
coeff    0.040491                  0.189495     2.38318

       Is.Non.Annual._Imputed  SingleGradeTripFlag_Imputed  \
coeff               -2.837825                     0.874364

       SPR.New.Existing_Imputed
coeff                 -1.512815

AIC 8573.747004282925
```

Step 3- Evaluate the model: Evaluate the model's performance on the validation data. We have obtained an accuracy of **80.36%.** The Odds ratio obtained for the prediction accuracy of the validation data is 15.02.

```
Confusion Matrix (Accuracy 0.8036)

       Prediction
Actual   0   1
     0 263 103
     1  85 506
```

	Estimate	SE	LCB	UCB	p-value
Odds ratio	15.200		10.998	21.007	0.000
Log odds ratio	2.721	0.165	2.398	3.045	0.000
Risk ratio	4.996		4.062	6.145	0.000
Log risk ratio	1.609	0.106	1.402	1.816	0.000

c) Comparing the predictive accuracy of all the models we used:

a. KNN model – with a K value of 9 we are achieving an accuracy of 79.51%

b. Decision Tree Classification - 78.89%

c. Random Forest Classification – 79.41%

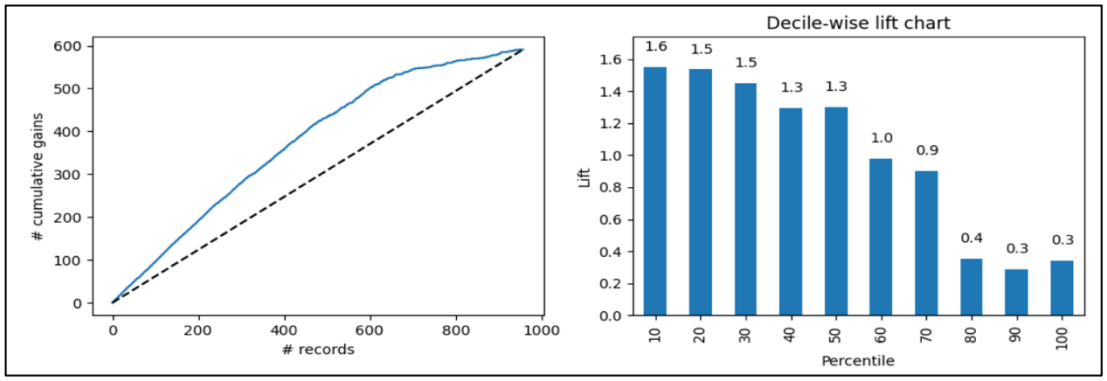
d. Gradient Booster Classification – 79.62%

e. Logit Regression Model – 80.36%

Based on the reported accuracy values, the Logit Regression Model has the highest accuracy of 80.36% on the validation data. We have obtained this accuracy by considering 6 prediction variables.

We have considered 9 prediction variables for the other classification models. Accuracy obtained for the Gradient Booster Classification is 79.62%, then we have Random Forest Classification with 79.41% accuracy, KNN model with a K value of 9 with 79.51% accuracy, and Decision Tree Classification with 78.89% accuracy.

Please find below cumulative gains and Decile-wise lift chart.



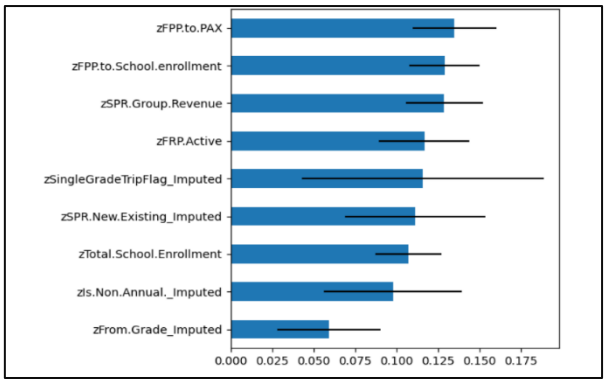
That said, it's important to note that accuracy alone may not be the only factor to consider when selecting a model. Additionally, it's crucial to assess the performance of the models on various evaluation metrics and validation data before selecting the final model for deployment.

d) Based on the feature importance analysis, the predictors that have the highest importance in predicting the outcome variable are:

- **FPP.to.PAX** : The ratio of FPP to total PAX on the trip.
- **FPP.to.School.enrollment** : The ratio of FPP to school enrollment.

- **SPR.Group.Revenue** : The total amount paid for all the participants to go on to the program from that group.
- **FRP.Active** : FRP is the full refund program. This the no. of FPP's on the trip who bought trip cancellation insurance.
- **SingleGradeTripFlag** : Indicator for the trip taken by a group comprising students from the same grade.
- **SPR.New.Existing** : Existing means the group has travelled with STC before – most often the year before. NEW with few exceptions, means that the school has never travelled with STC before.

	feature	importance	std
5	zFrom.Grade_Imputed	0.059246	0.031059
6	zIs.Non.Annual._Imputed	0.097837	0.041570
1	zTotal.School.Enrollment	0.107227	0.019974
8	zSPR.New.Existing_Imputed	0.111075	0.042315
7	zSingleGradeTripFlag_Imputed	0.115630	0.072708
0	zFRP.Active	0.116674	0.027122
2	zSPR.Group.Revenue	0.128700	0.023030
3	zFPP.to.School.enrollment	0.128869	0.021035
4	zFPP.to.PAX	0.134742	0.025168



e) Interpret the meaning of the logit coefficients for at least two influential variables.

When interpreting the logit coefficients for influential variables in a logistic regression model, we typically look at the sign of the coefficient (positive or negative) and the magnitude of the coefficient. The sign of the coefficient indicates the direction of the relationship between the predictor variable and the outcome variable (i.e., whether the variable increases or decreases the likelihood of the outcome), while the magnitude of the coefficient indicates the strength of the relationship.

In this case the coefficients are shown below:

```
intercept  -1.824349468222632
       FRP.Active  FPP.to.School.enrollment  FPP.to.PAX  \
coeff    0.040491                  0.189495     2.38318

       Is.Non.Annual._Imputed  SingleGradeTripFlag_Imputed  \
coeff               -2.837825                     0.874364

       SPR.New.Existing_Imputed
coeff                 -1.512815

AIC 8573.747004282925
```

Interpretation:

a. For FPP.to.PAX: The coefficient is 2.38, which means that a one-unit increase in FPP.to.PAX increases the log-odds of the outcome by 2.38 units, provided all other variables are held constant. In other words, an increase in FPP.to.PAX is associated with a higher likelihood of the outcome.

b. For Is.Non.Annual.: The coefficient is -2.84. This variable is a binary variable, as 0 represents "annual" and 1 represents "non-annual." When Is.Non.Annual holds the value as 1, it decreases the log-odds of the outcome by 2.84 units, provided all other variables are held constant. Therefore, this coefficient suggests that customers who have a non-annual frequency of purchase are less likely to exhibit the outcome of interest compared to customers who purchase annually.

f) Predicting Odds for the logit model

As mentioned earlier, we have obtained an accuracy of 80.36%. The Odds ratio obtained for the prediction accuracy of the validation data is 15.02.

```
Confusion Matrix (Accuracy 0.8036)

       Prediction
Actual   0   1
     0 263 103
     1  85 506
```

	Estimate	SE	LCB	UCB	p-value
Odds ratio	15.200		10.998	21.007	0.000
Log odds ratio	2.721	0.165	2.398	3.045	0.000
Risk ratio	4.996		4.062	6.145	0.000
Log risk ratio	1.609	0.106	1.402	1.816	0.000

We have introduced a new customer record for checking the prediction accuracy of the logit model.

FRP.Active	FPP.to.School.enrollment	FPP.to.PAX	Is.Non.Annual._Imputed	SingleGradeTripFlag_Imputed	SPR.New.Existing_Imputed
12	1	0.9	0	1.0	1

```
Confusion Matrix (Accuracy 1.0000)

       Prediction
Actual 0
     0 1
```

	Estimate	SE	LCB	UCB	p-value
Odds ratio	2.000		0.011	357.353	0.793
Log odds ratio	0.693	2.646	-4.492	5.879	0.793
Risk ratio	1.500		0.075	29.945	0.791
Log risk ratio	0.405	1.528	-2.588	3.399	0.791

The prediction accuracy we have obtained for the new data is 100% and the odds ratio is 2.