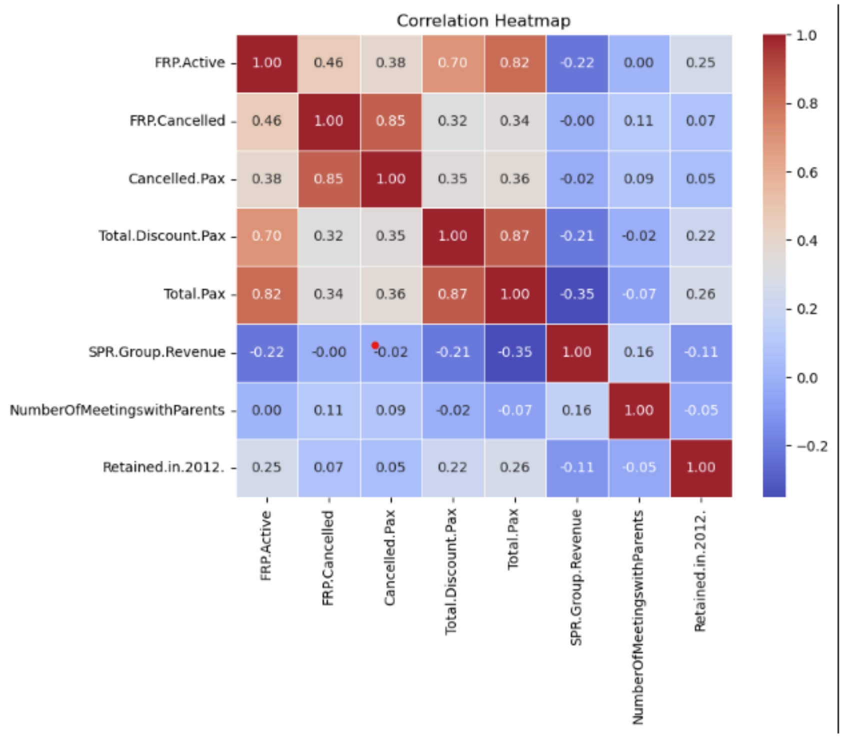
BAN 620 – DATA MINING
HW4 – Group 6

Vatyam Sarath Kumar
Abhi Ram Sai Lakkavarapu
Neeharika Vijjada

STC is a travel agency company that manages the trips related to educational. Given data is about the customer list of STC and this data belongs to 2010-11. The interest variable in this data set is Retained.in.2012. which give the information whether that group who has already booked a trip with STC, booked again with them in the year 2012.

This dataset has 2392 records and 56 columns and has multiple datatypes.
Data cleaning has been done prior to the analysis which involved in changing the datatypes of multiple variables and dropping the variables from the data, which is not having any significance with the prediction whether the customer booked with STC again or not.

A correlation heat map has been developed to explore the relation between numerical variables and the outcome variable "Retained.in.2012."

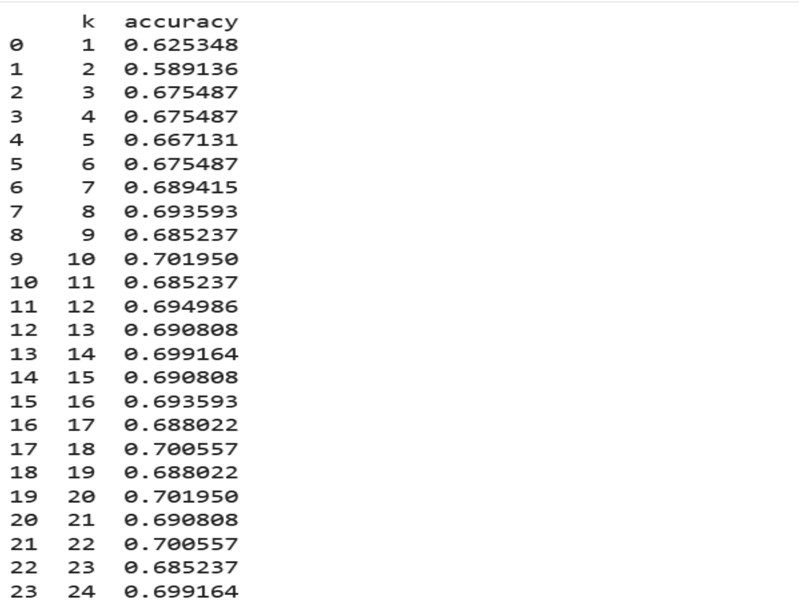


	FRP.Active	FRP.Cancelled	Cancelled.Pax	Total.Discount.Pax	Total.Pax	SPR.Group.Revenue	NumberOfMeetingswithParents	Retained.in.2012.
FRP.Active	1.00	0.46	0.38	0.70	0.82	-0.22	0.00	0.25
FRP.Cancelled	0.46	1.00	0.85	0.32	0.34	-0.00	0.11	0.07
Cancelled.Pax	0.38	0.85	1.00	0.35	0.36	-0.02	0.09	0.05
Total.Discount.Pax	0.70	0.32	0.35	1.00	0.87	-0.21	-0.02	0.22
Total.Pax	0.82	0.34	0.36	0.87	1.00	-0.35	-0.07	0.26
SPR.Group.Revenue	-0.22	-0.00	-0.02	-0.21	-0.35	1.00	0.16	-0.11
NumberOfMeetingswithParents	0.00	0.11	0.09	-0.02	-0.07	0.16	1.00	-0.05
Retained.in.2012.	0.25	0.07	0.05	0.22	0.26	-0.11	-0.05	1.00

From the above heat map, it is evident that there were few variables which are having a strong correlation with other variables in the dataset.
Generally, it is common that if there are more number of passengers then there are high chances that more customers would opt in the insurance. As it's having a linear relation.

Upon performing the exploratory data analysis, multiple models were created with this data. Among the models made, the first model was KNN, which is used to predict the retention rate of the previous customers. Since, KNN model was used, it is essential that all the variables has to be normalized, but there are few categorical variables which has been converted to numerical by creating dummies with the variables which were considered, expecting some significance with the target variable.

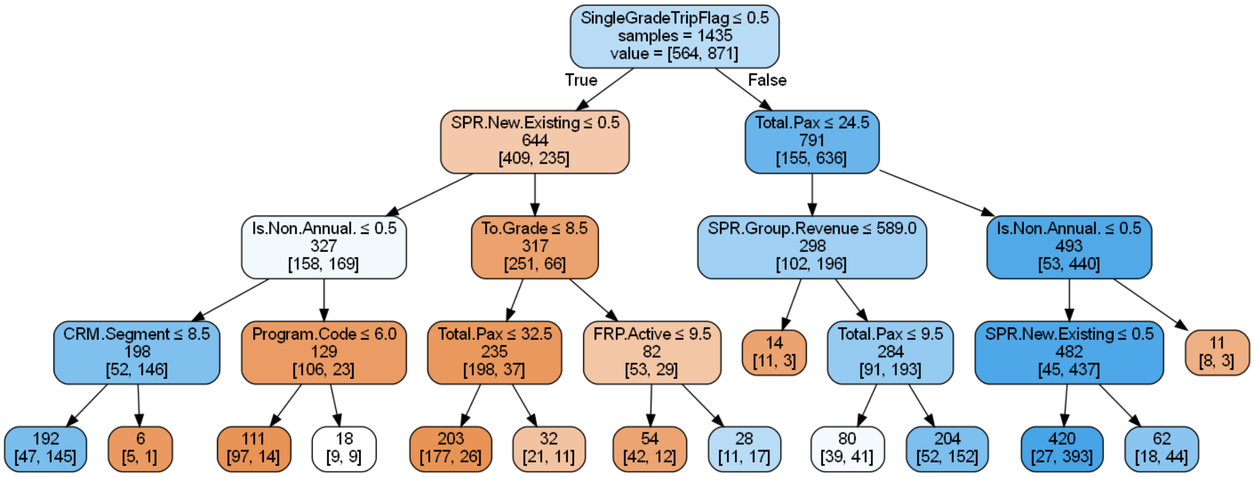
Initially, we have considered the nearest neighbors count of 3 in the model, but upon finding the accuracies with multiple K-values, it's seen that at K=10, we can observe a maximum accuracy of 70.19%. So, this k-value has been considered and model has been built on this characteristic. This model was then used to predict the validation data.



```
        k  accuracy
0       1  0.625348
1       2  0.589136
2       3  0.675487
3       4  0.675487
4       5  0.667131
5       6  0.675487
6       7  0.689415
7       8  0.693593
8       9  0.685237
9      10  0.701950
10     11  0.685237
11     12  0.694986
12     13  0.690808
13     14  0.699164
14     15  0.690808
15     16  0.693593
16     17  0.688022
17     18  0.700557
18     19  0.688022
19     20  0.701950
20     21  0.690808
21     22  0.700557
22     23  0.685237
23     24  0.699164
```

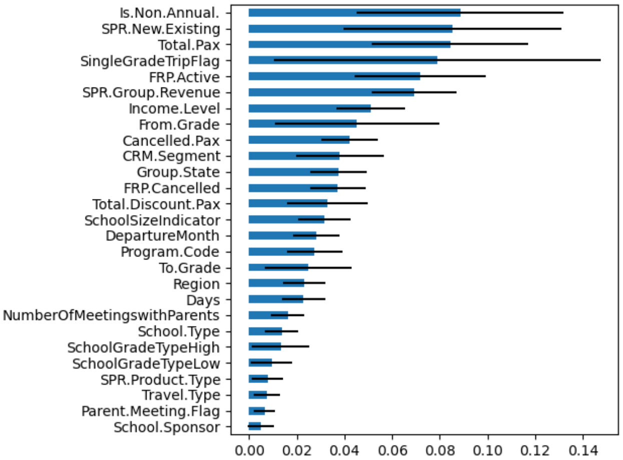
Next, a classification tree model was developed by categorizing the variables into predictors and outcome variables. Upon categorizing, this data has been split into test and validation data in the ratio 60:40 respectively. A decision tree has been developed which has 27 features in it.

Since the characteristics has been selected randomly, in order to select the better characteristics we have used grid search method and came up with characteristics which gave an accuracy of 79.4%



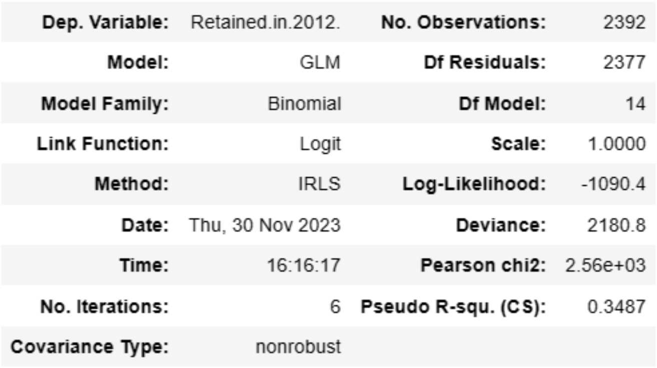
One of the rules that can be formed from the tree is if all the passengers/students on the trip is of same grade and if total passengers are less than 24 and if it’s a non annual trip then the retention chances are very high.

Later, Random forests were used to determine the important variables from the dataset which are significant with the target variable.



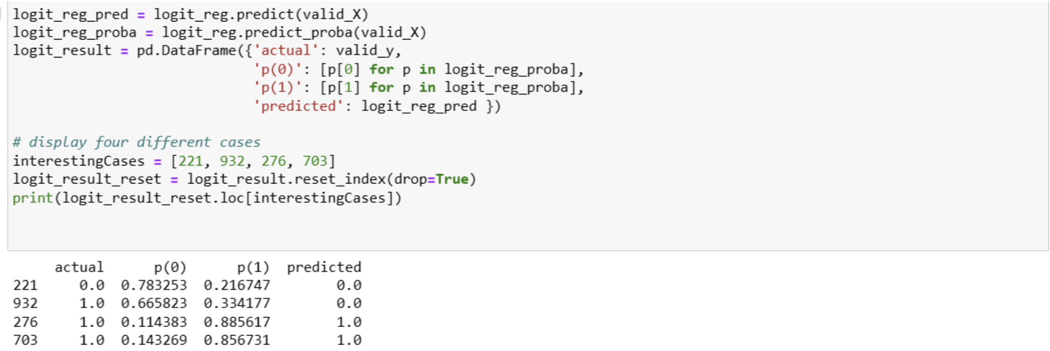
Above chart, will give an understanding of the importance of variables with respect to target variable ‘Retained.in2012.’

Similar prediction model was developed using Logistic regression.
All the significant predictors were considered in this model along with the outcome variable. A logistic summary was derived from the initial analysis



Dep. Variable:	Retained.in.2012.	**No. Observations:**	2392
Model:	GLM	**Df Residuals:**	2377
Model Family:	Binomial	**Df Model:**	14
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1090.4
Date:	Thu, 30 Nov 2023	**Deviance:**	2180.8
Time:	16:16:17	**Pearson chi2:**	2.56e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.3487
Covariance Type:	nonrobust		

Also coefficients of each variable with respect to the target variable is also obtained as part of this analysis.
Data has been again partitioned into test and validation data, and then passed into the model for training. For the naive model, AIC score was 1212.61. Then, validation data was passed into the model and few random datapoints are tested for validation.



```
logit_reg_pred = logit_reg.predict(valid_X)
logit_reg_proba = logit_reg.predict_proba(valid_X)
logit_result = pd.DataFrame({'actual': valid_y,
                             'p(0)': [p[0] for p in logit_reg_proba],
                             'p(1)': [p[1] for p in logit_reg_proba],
                             'predicted': logit_reg_pred })

# display four different cases
interestingCases = [221, 932, 276, 703]
logit_result_reset = logit_result.reset_index(drop=True)
print(logit_result_reset.loc[interestingCases])
```

```
     actual      p(0)      p(1)  predicted
221     0.0  0.783253  0.216747        0.0
932     1.0  0.665823  0.334177        0.0
276     1.0  0.114383  0.885617        1.0
703     1.0  0.143269  0.856731        1.0
```

From classification summary, the accuracy of validation data is 79.94%

We have prepared three models namely – KNN, Classification trees, Logit. Each model has different accuracy as follows

Model	Accuracy
KNN Model	70.19%
Classification Tree model	79.40%
Logit Model	79.94%

By comparing three models, it's recommended to use Logit model as it's having better accuracy for the validation data compared to other two models.

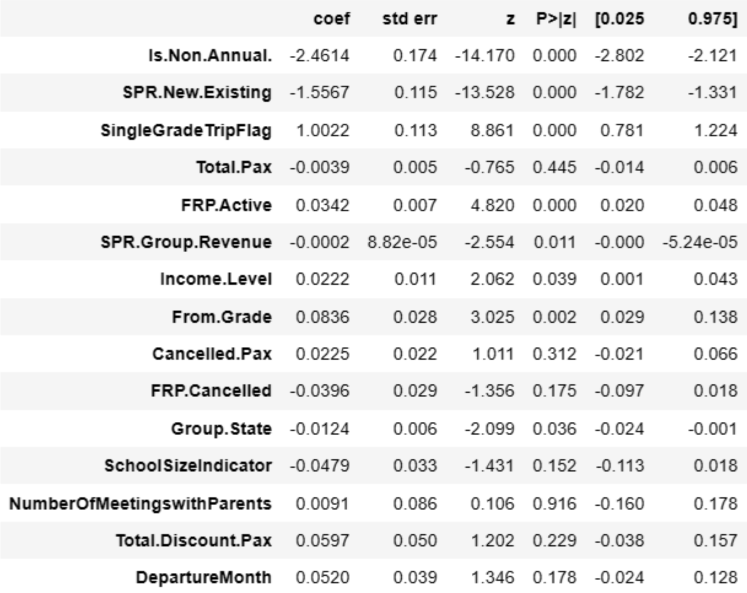
As mentioned earlier, the original dataset has 56 columns, but all the variables in those dataset are not useful for our analysis. So we have dropped few columns and considered below variables for the model development

'Program.Code', 'From.Grade', 'To.Grade', 'Group.State',
'Is.Non.Annual.', 'Days', 'Travel.Type', 'FRP.Active', 'FRP.Cancelled',
'Cancelled.Pax', 'Total.Discount.Pax', 'Region', 'CRM.Segment',
'School.Type', 'Parent.Meeting.Flag', 'Income.Level', 'School.Sponsor',
'SPR.Product.Type', 'SPR.New.Existing', 'Total.Pax',
'SPR.Group.Revenue', 'NumberOfMeetingswithParents',
'SchoolGradeTypeLow', 'SchoolGradeTypeHigh', 'DepartureMonth',
'SingleGradeTripFlag', 'SchoolSizeIndicator'

For the best model(logit model), only few predictors were used as per the Random forest variable importance plot

'Is.Non.Annual.','SPR.New.Existing', 'SingleGradeTripFlag', 'Total.Pax', 'FRP.Active',
'SPR.Group.Revenue','Income.Level', 'From.Grade', 'Cancelled.Pax', 'FRP.Cancelled',
'Group.State','SchoolSizeIndicator', 'NumberOfMeetingswithParents', 'Total.Discount.Pax',
'DepartureMonth'

In logistic regression summary, coefficients were derived for the predictor variables with the target variable



	coef	std err	z	P>|z|	[0.025	0.975]
Is.Non.Annual.	-2.4614	0.174	-14.170	0.000	-2.802	-2.121
SPR.New.Existing	-1.5567	0.115	-13.528	0.000	-1.782	-1.331
SingleGradeTripFlag	1.0022	0.113	8.861	0.000	0.781	1.224
Total.Pax	-0.0039	0.005	-0.765	0.445	-0.014	0.006
FRP.Active	0.0342	0.007	4.820	0.000	0.020	0.048
SPR.Group.Revenue	-0.0002	8.82e-05	-2.554	0.011	-0.000	-5.24e-05
Income.Level	0.0222	0.011	2.062	0.039	0.001	0.043
From.Grade	0.0836	0.028	3.025	0.002	0.029	0.138
Cancelled.Pax	0.0225	0.022	1.011	0.312	-0.021	0.066
FRP.Cancelled	-0.0396	0.029	-1.356	0.175	-0.097	0.018
Group.State	-0.0124	0.006	-2.099	0.036	-0.024	-0.001
SchoolSizeIndicator	-0.0479	0.033	-1.431	0.152	-0.113	0.018
NumberOfMeetingswithParents	0.0091	0.086	0.106	0.916	-0.160	0.178
Total.Discount.Pax	0.0597	0.050	1.202	0.229	-0.038	0.157
DepartureMonth	0.0520	0.039	1.346	0.178	-0.024	0.128

From the above table, for variable Is.Non.Annual. there is a -2.46 coefficient with the target variable retention.in.2012.
This negative sign indicates a inverse relationship and the magnitude gives strength of relation ship. In this case If the trip is an Annual Trip(Is.Non.Annual =0) then there are very less chances that they will again take the trip in 2012 as , if 'Is.Non.Annual' increases by 1 unit, the log odds of the event 'Retained.in.2012.' occurring will decrease by 2.4614

Similarly, for the variable SingleGradeTripFlag, the coef is +1.002 with outcome variable, As it's having a positive sign this states that the variables has a linear relation as coefficient is almost 1. So, if there are same grade people travelling on a trip then there is a high chance (retention.in.2012 =1) that will take the trip. If SingleGradeTripFlag increases by 1 unit, the log oddsd of the outcome event will increase by 1 time.

To check the odds of the logit model, a new data has been passed to the model.

'Is.Non.Annual.':1.0,'SPR.New.Existing':1.0,'SingleGradeTripFlag':1,'Total.Pax':55.0,'FRP.Active':6.0,'SPR.Group.Revenue':1885.0,'Income.Level':6,'From.Grade':7,'Cancelled.Pax':7.0,'FRP.Cancelled':2.0,'Group.State':20,'SchoolSizeIndicator':1,'NumberOfMeetingswithParents':2,'Total.Discount.Pax':2,'DepartureMonth':4

The odds of booking trip with STC for the academic year 2013 is 10.45 with the given new input data.