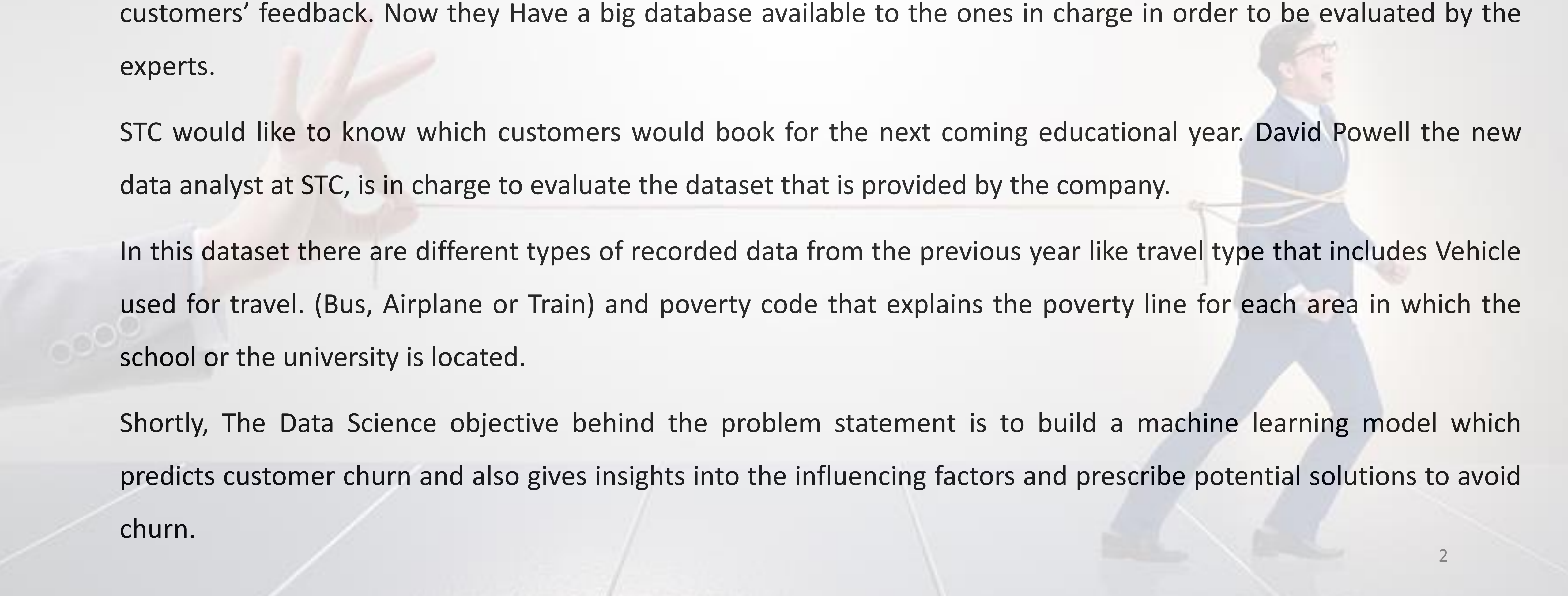
Background

Today, with the advancement of technology, the business world has become very competitive and challenging to the extent that as opposed to the period when Henry Ford claimed that every customer can choose “any color so long as it is black”, there are many options and brands available to the customers these days that make customer satisfaction and retention policies and strategies more important than ever.

Customer loyalty and customer retention are corelated. Customer loyalty is a measure of a customer’s likeliness to choose a company or brand again. It is the result of customer satisfaction, positive customer experiences, and the overall value of the goods or services a customer is given by a company.

To do so, we need to identify the needs and expectations of our customers and provide services appropriate to those features. On the other hand, weakness must be identified so that companies can make necessary actions to rectify them.



Problem Statement

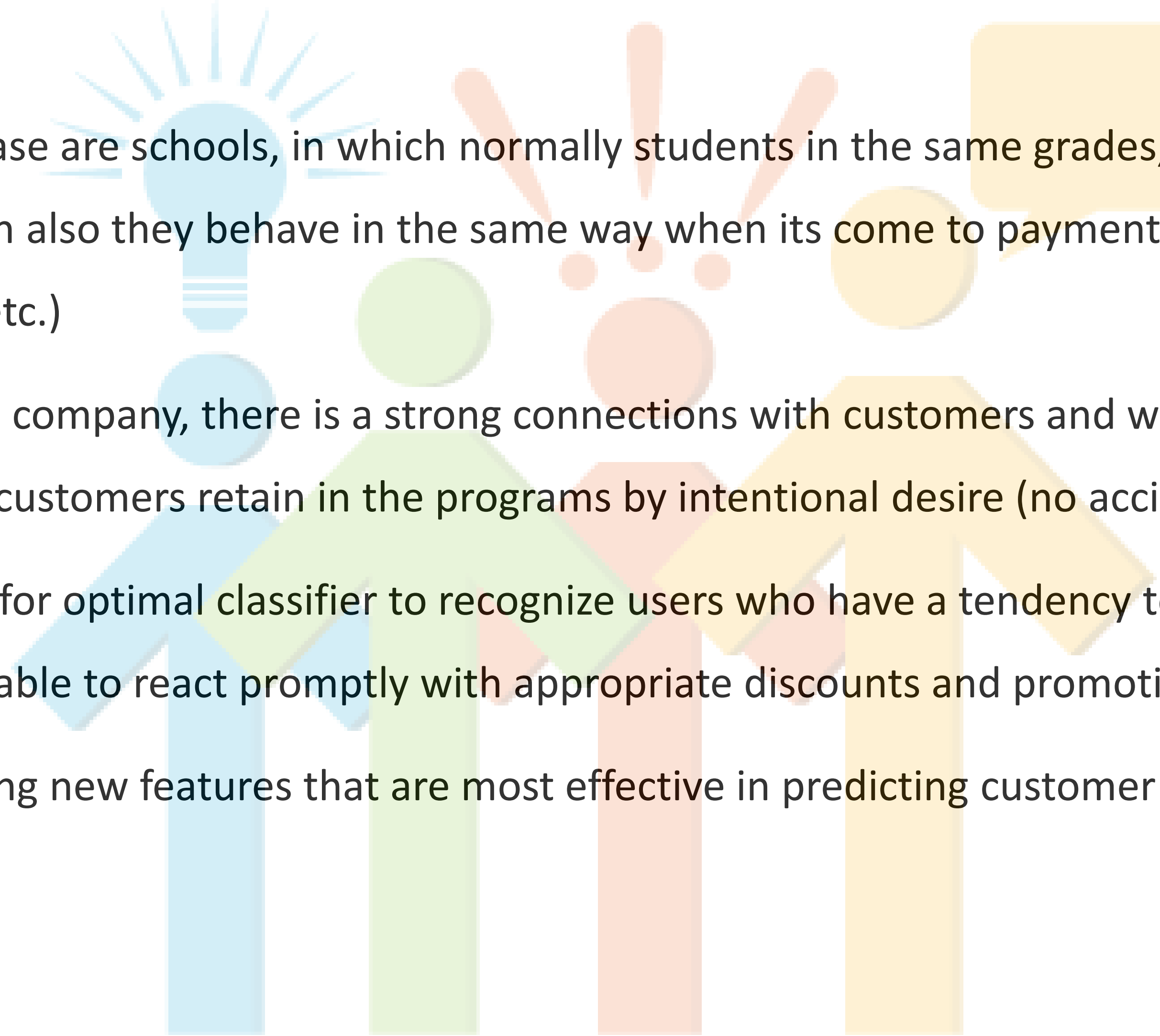
STC is one of the premium providers of different types of travels such as historical, science and educational. For high school and university students throughout the world. They collect verity types of data related to the travels and customers' feedback. Now they Have a big database available to the ones in charge in order to be evaluated by the experts.

STC would like to know which customers would book for the next coming educational year. David Powell the new data analyst at STC, is in charge to evaluate the dataset that is provided by the company.

In this dataset there are different types of recorded data from the previous year like travel type that includes Vehicle used for travel. (Bus, Airplane or Train) and poverty code that explains the poverty line for each area in which the school or the university is located.

Shortly, The Data Science objective behind the problem statement is to build a machine learning model which predicts customer churn and also gives insights into the influencing factors and prescribe potential solutions to avoid churn.

Assumptions



- Target society in this case are schools, in which normally students in the same grades, tend to travel with their friends in any condition also they behave in the same way when its come to payment related issues (such as refund condition and etc.)
- We assume that in this company, there is a strong connections with customers and we just want to ensure that they retain. And also, customers retain in the programs by intentional desire (no accidental customer churn)
- Finally, we are looking for optimal classifier to recognize users who have a tendency to churn in the near future, so company will become able to react promptly with appropriate discounts and promotions
- We aimed for identifying new features that are most effective in predicting customer churn .

Agenda

A brief Look into the Dataset

Initial Steps

Exploratory Data Analysis

Feature Engineering (Feature Selection)

Preprocessing the Dataset

Applying Classification Algorithm

Choosing the Best Classification Algorithm

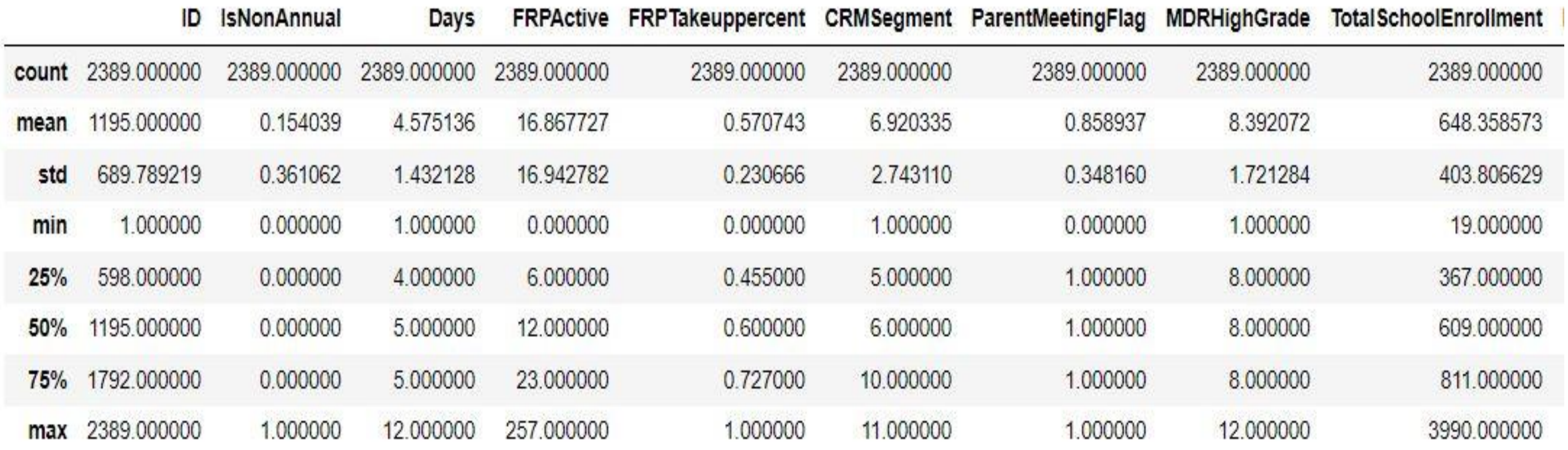
Algorithm Improvement

Forecasting the Following Year (Using Chosen Algorithm)

Conclusion and Recommendation

A brief Look into the Dataset

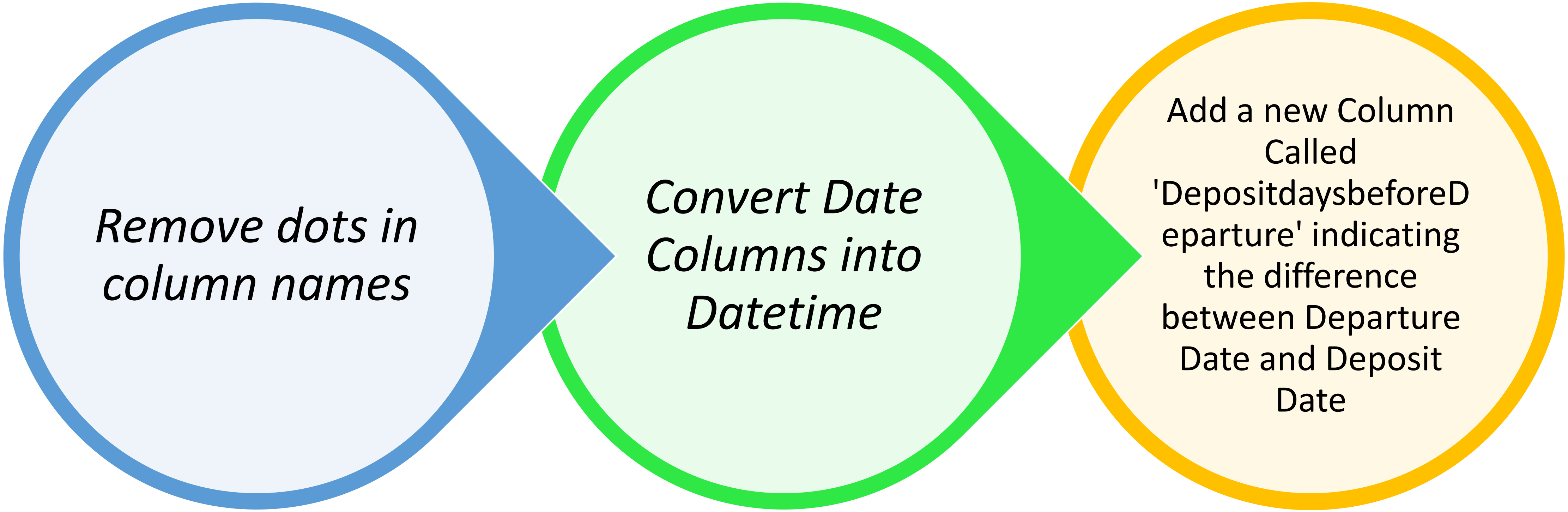
- *A dataset comprising 56 columns and 2389 rows.*
- *18 columns have NAN values*
- *The target value is imbalanced with a high proportion of Retained Customer to those who left in 2012*



	ID	IsNonAnnual	Days	FRPActive	FRPTakeuppercent	CRMSegment	ParentMeetingFlag	MDRHighGrade	TotalSchoolEnrollment
count	2389.000000	2389.000000	2389.000000	2389.000000	2389.000000	2389.000000	2389.000000	2389.000000	2389.000000
mean	1195.000000	0.154039	4.575136	16.867727	0.570743	6.920335	0.858937	8.392072	648.358573
std	689.789219	0.361062	1.432128	16.942782	0.230666	2.743110	0.348160	1.721284	403.806629
min	1.000000	0.000000	1.000000	0.000000	0.000000	1.000000	0.000000	1.000000	19.000000
25%	598.000000	0.000000	4.000000	6.000000	0.455000	5.000000	1.000000	8.000000	367.000000
50%	1195.000000	0.000000	5.000000	12.000000	0.600000	6.000000	1.000000	8.000000	609.000000
75%	1792.000000	0.000000	5.000000	23.000000	0.727000	10.000000	1.000000	8.000000	811.000000
max	2389.000000	1.000000	12.000000	257.000000	1.000000	11.000000	1.000000	12.000000	3990.000000

Initial Analysis

Before starting data visualization, some steps should be taken to make further analyses easier



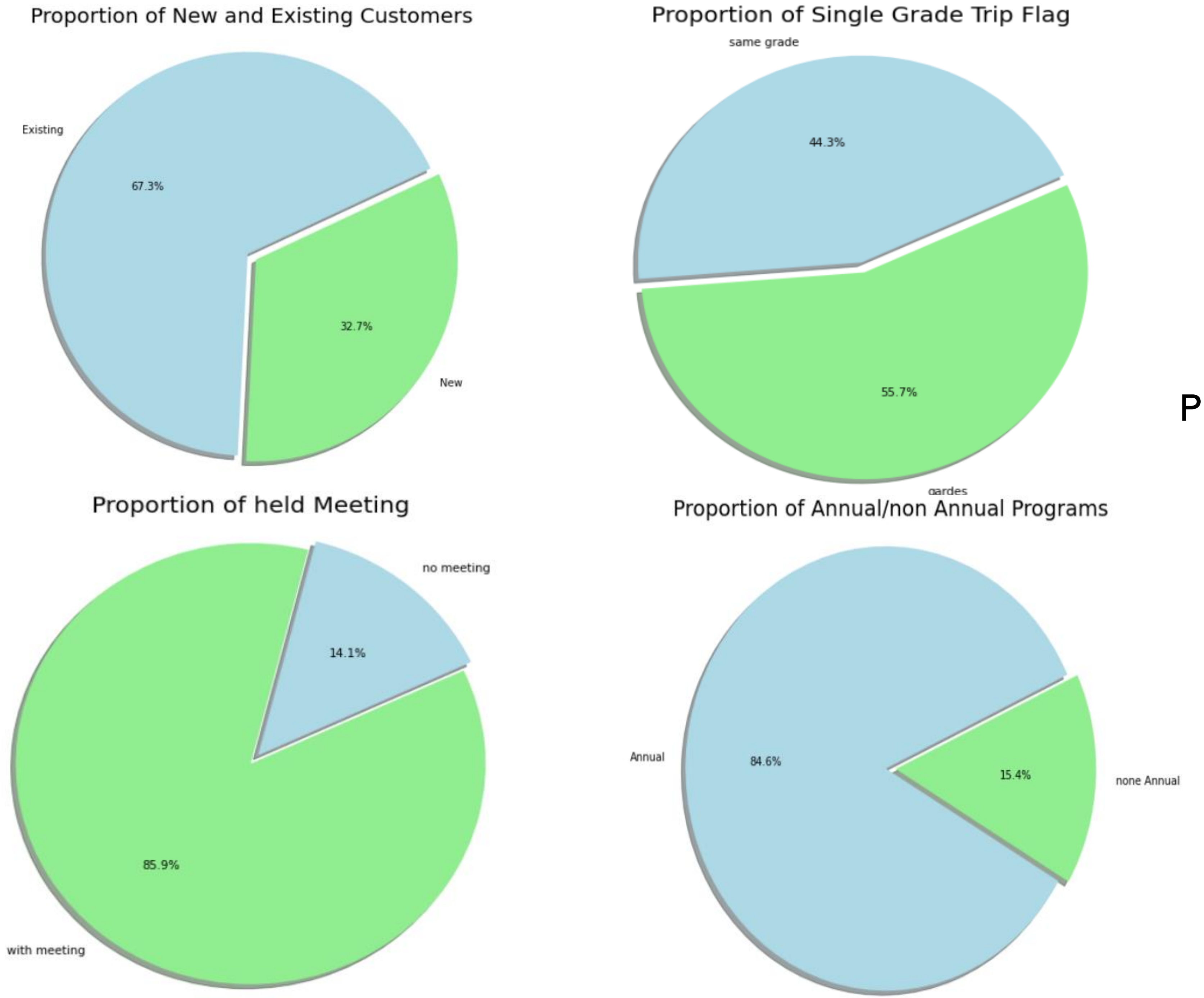
Remove dots in column names

Convert Date Columns into Datetime

Add a new Column Called 'DepositdaysbeforeDeparture' indicating the difference between Departure Date and Deposit Date

Exploratory Data Analysis

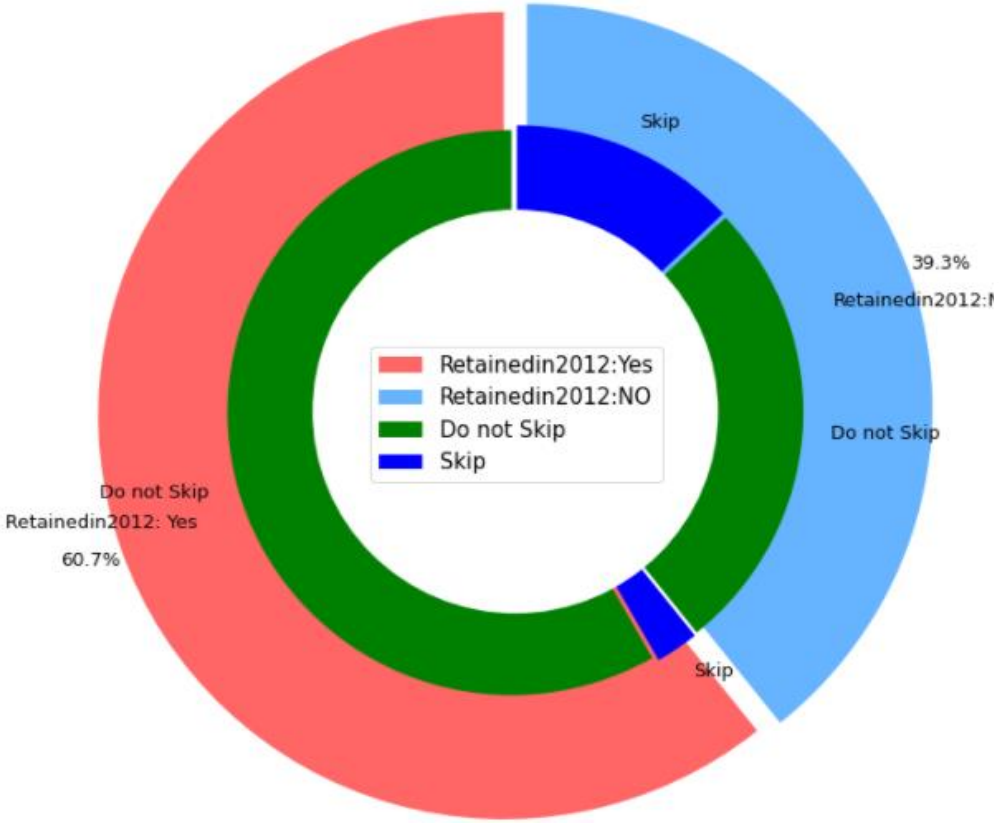
The following pages provides some graphs which were generated using Jupiter Notebook and Google Colab along with a brief explanation. This visualization narrows down to a conclusion which enables us to understand the main reason behind customer retention in 2012 as well as understanding factors affecting customer to make that decision.



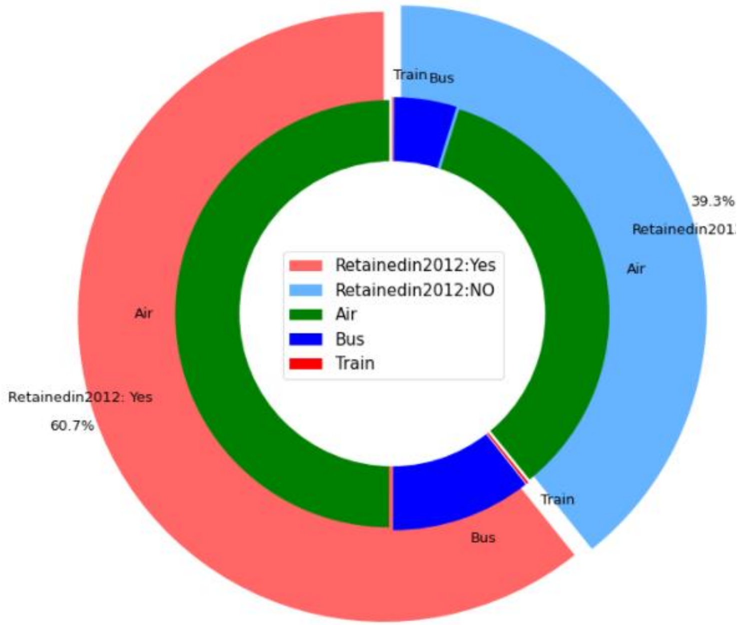
Proportion of four different binary features showing by pie charts

Comparing retention with different features

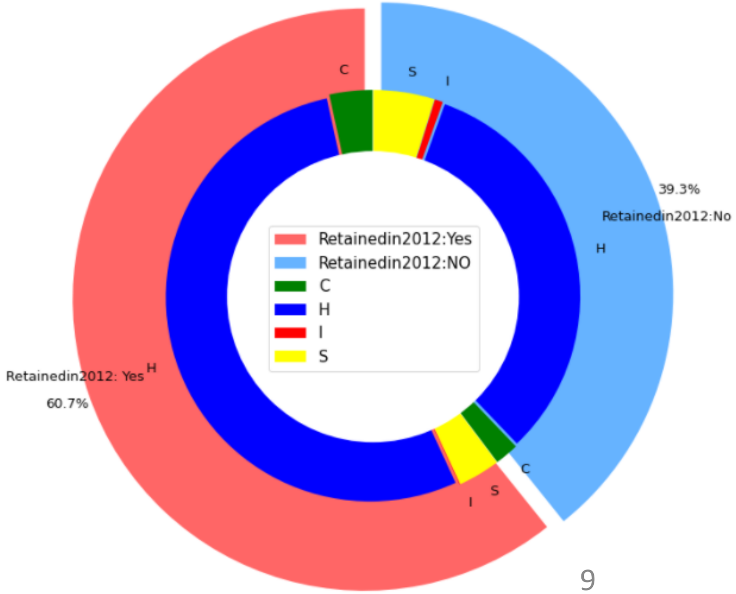
Retention Condition Distribution vs Travel Type



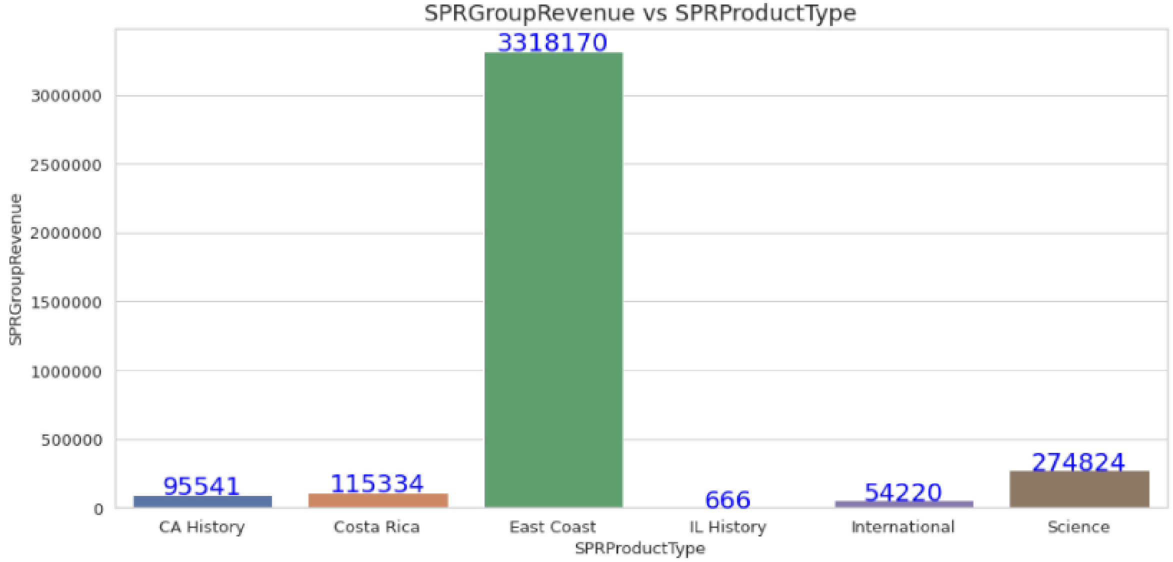
Retention Condition Distribution vs Travel Type

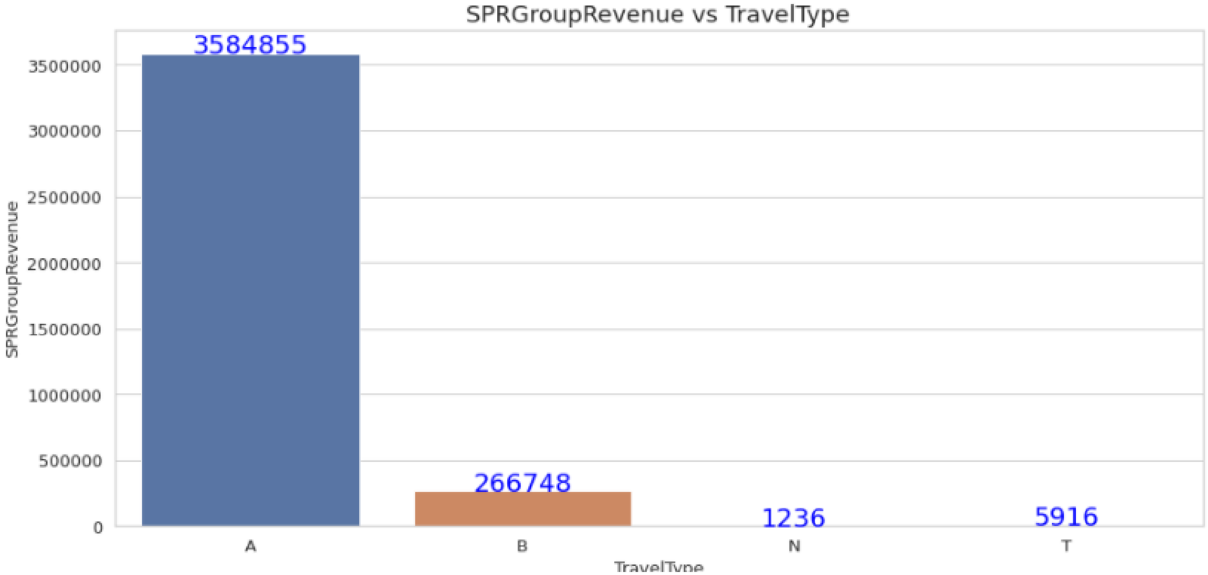


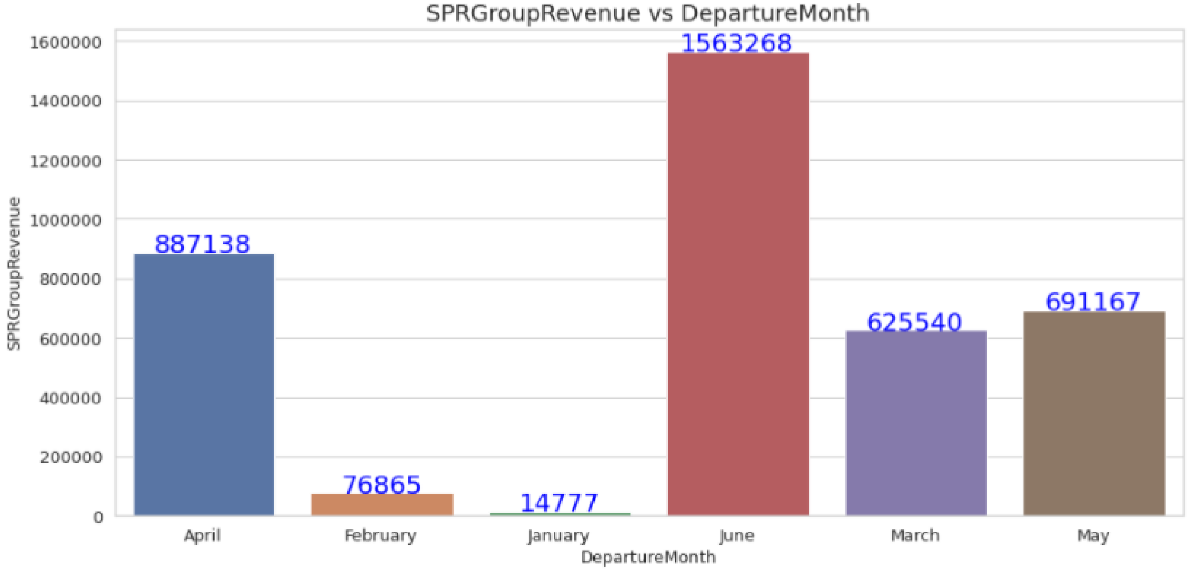
Retention Condition Distribution vs Major program code

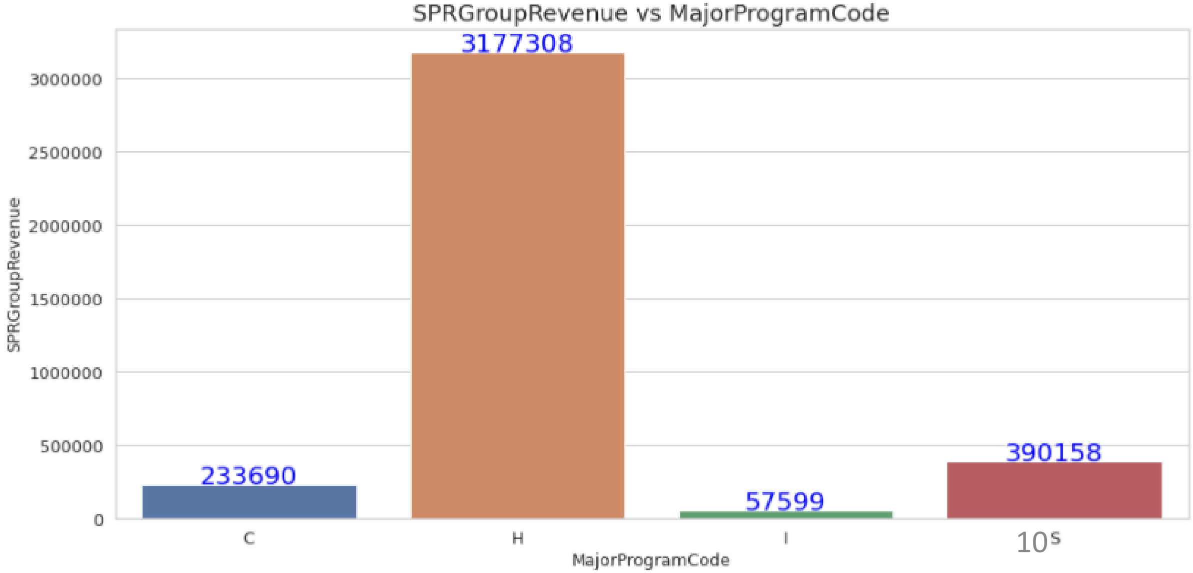


SPR Group Revenue through different features related to travel features to recognize customer needs before modeling customer retain rate

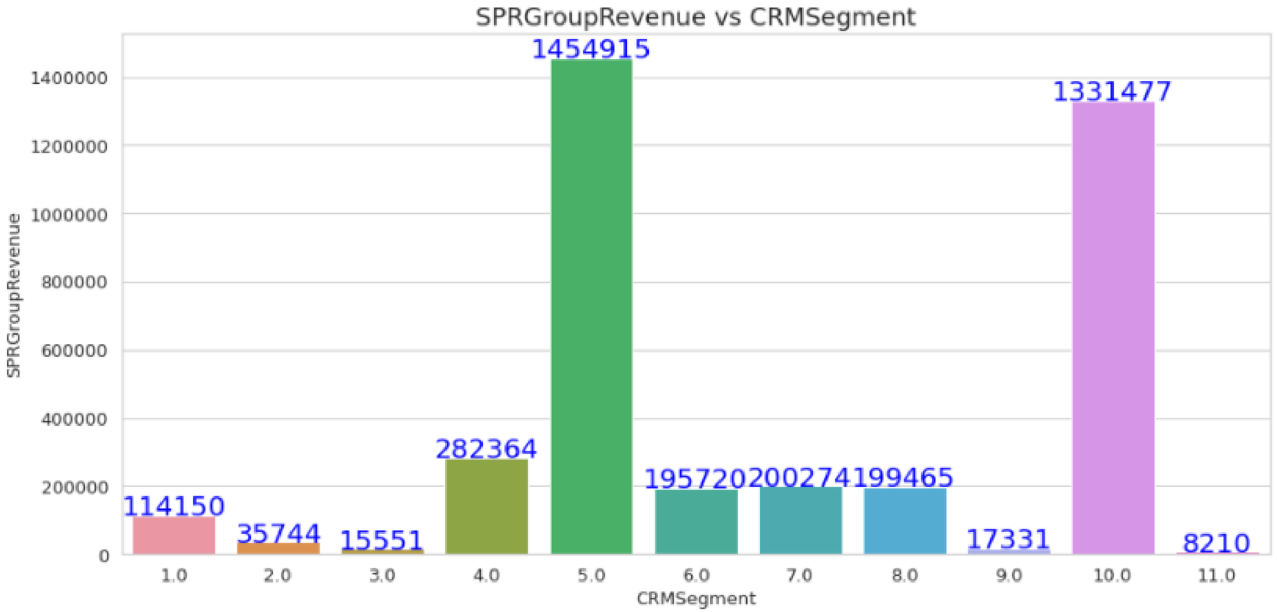


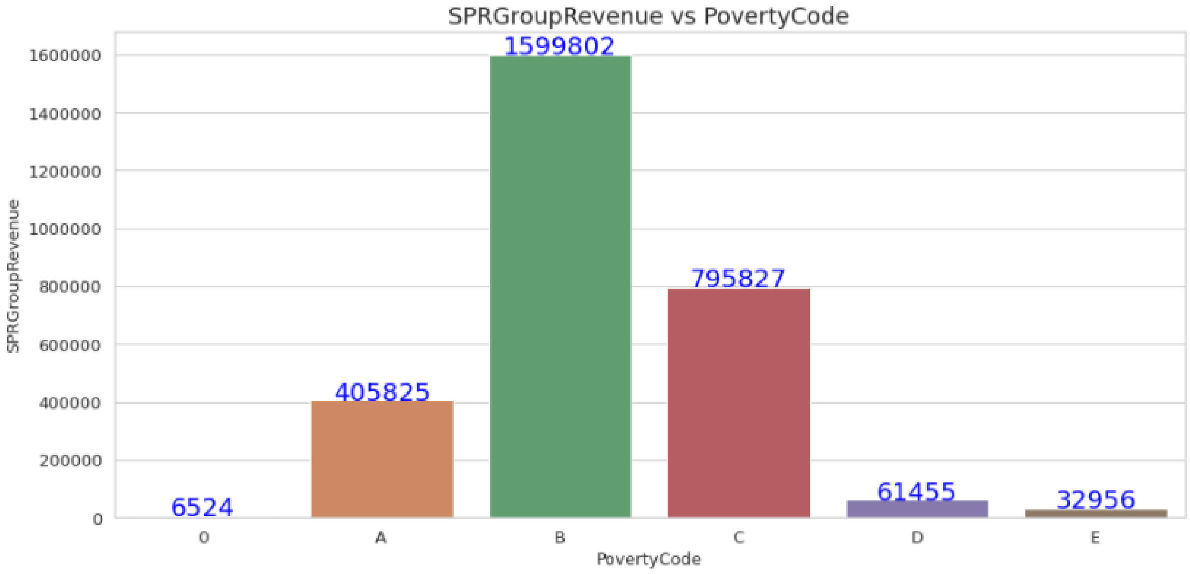


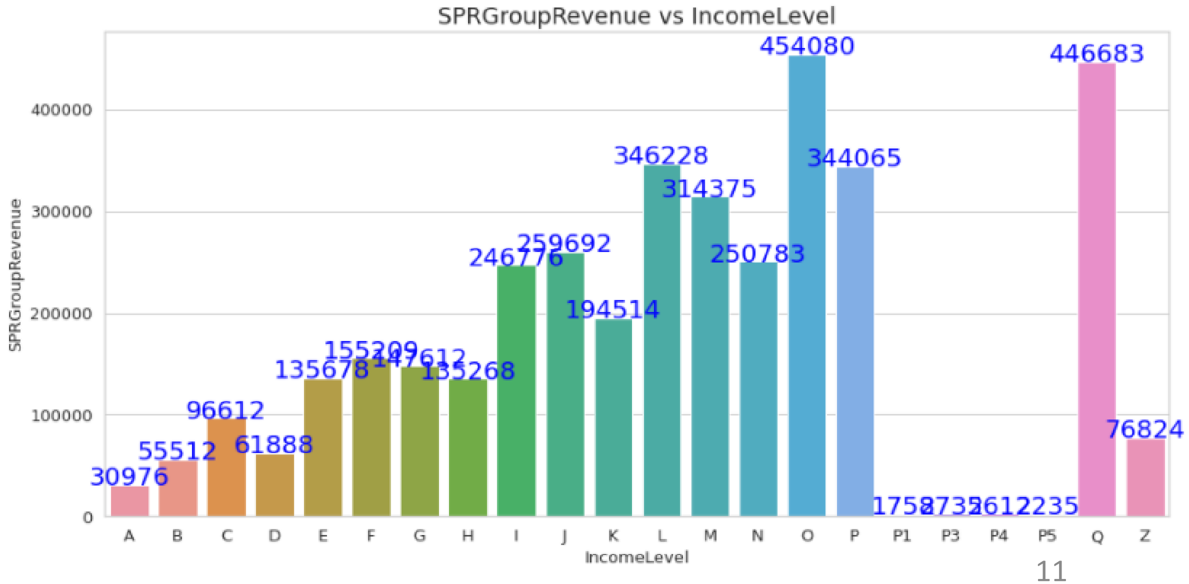




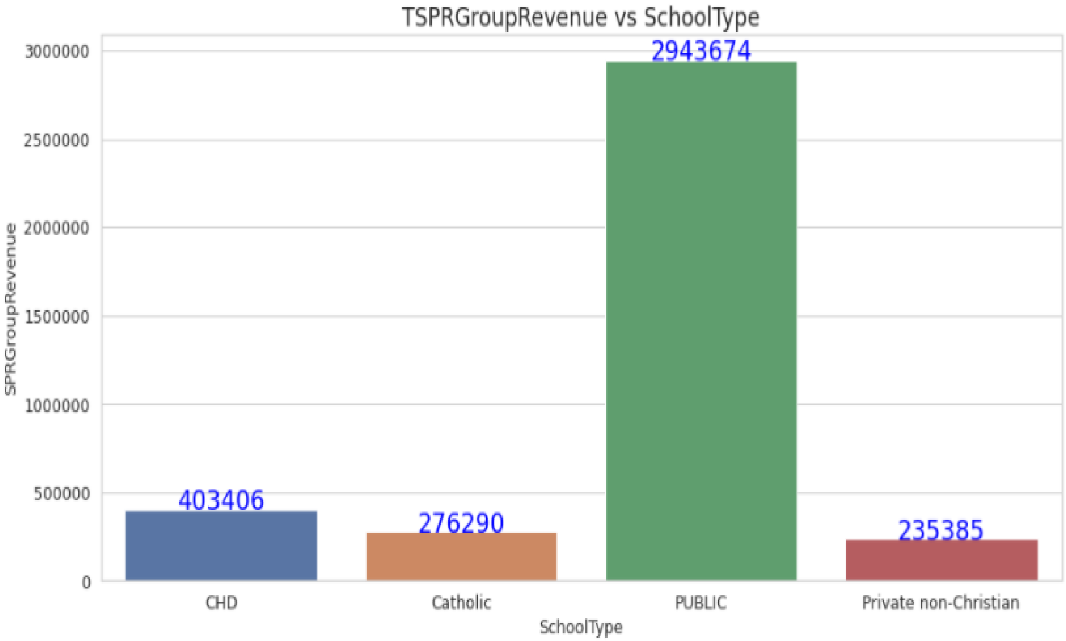
SPR Group Revenue through different direct features of the customer before modeling customer churn

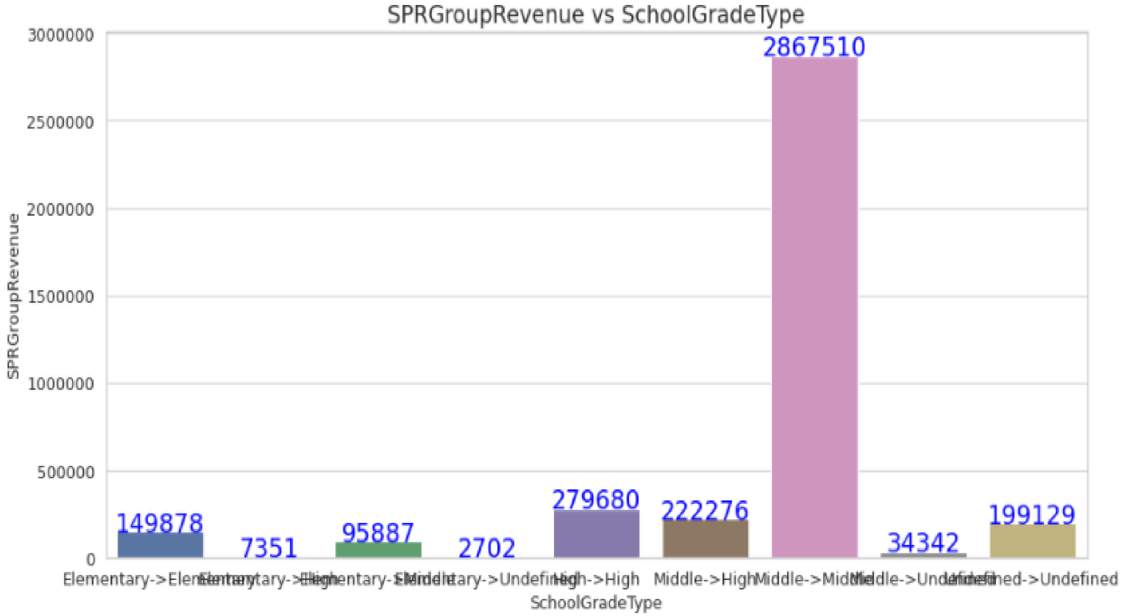


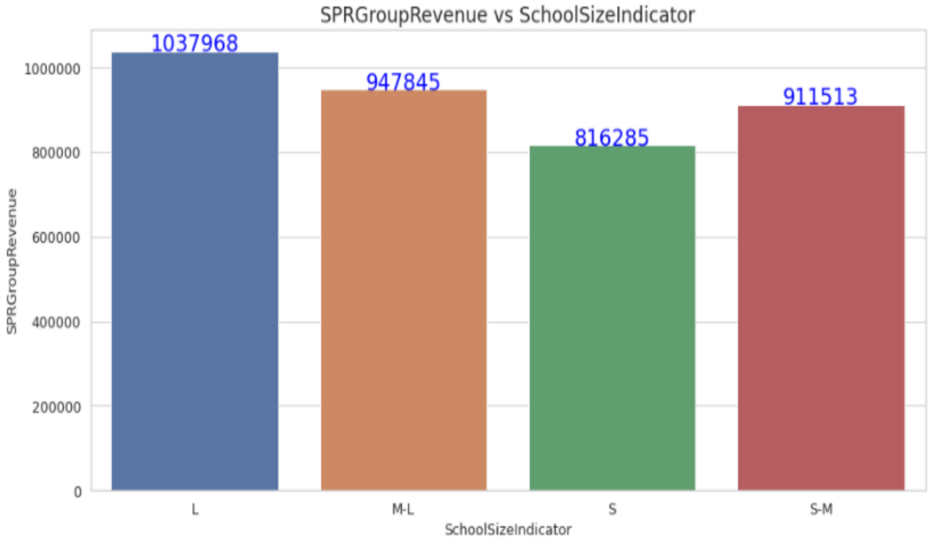




SPR Group Revenue through different school features before modeling customer churn

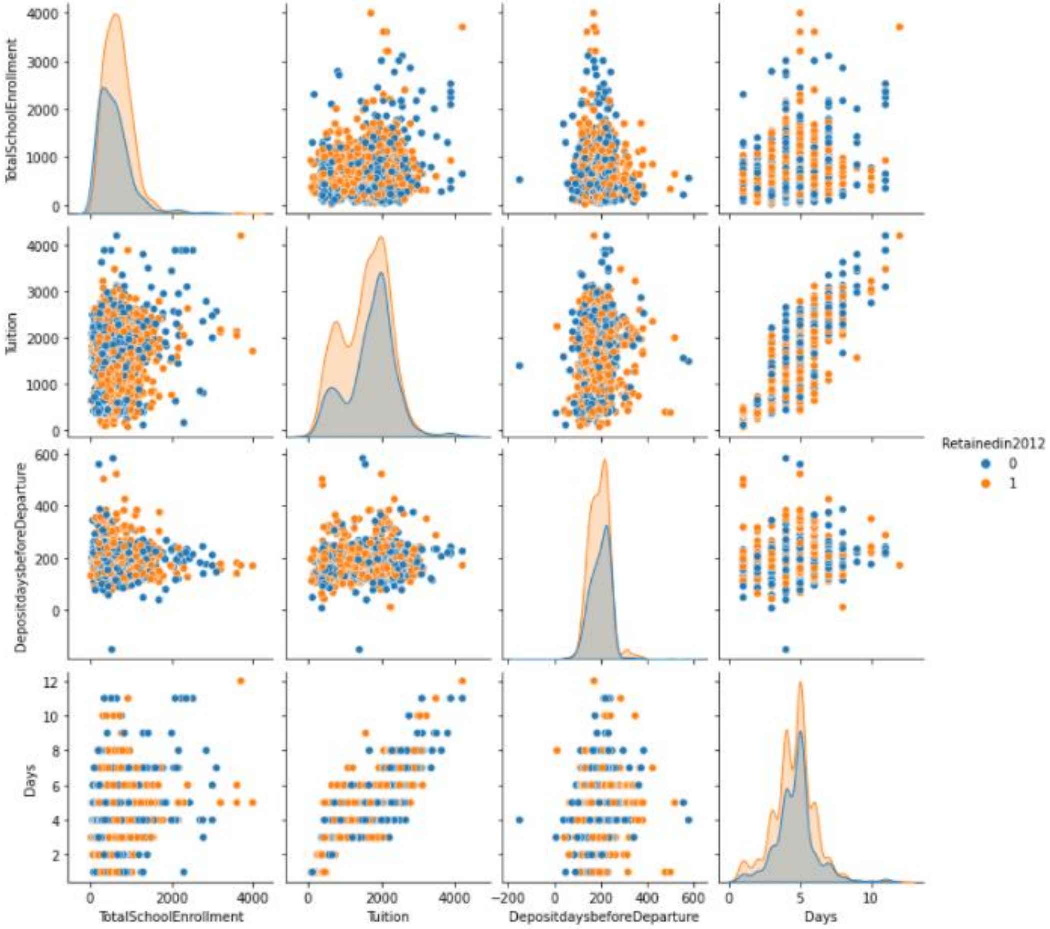






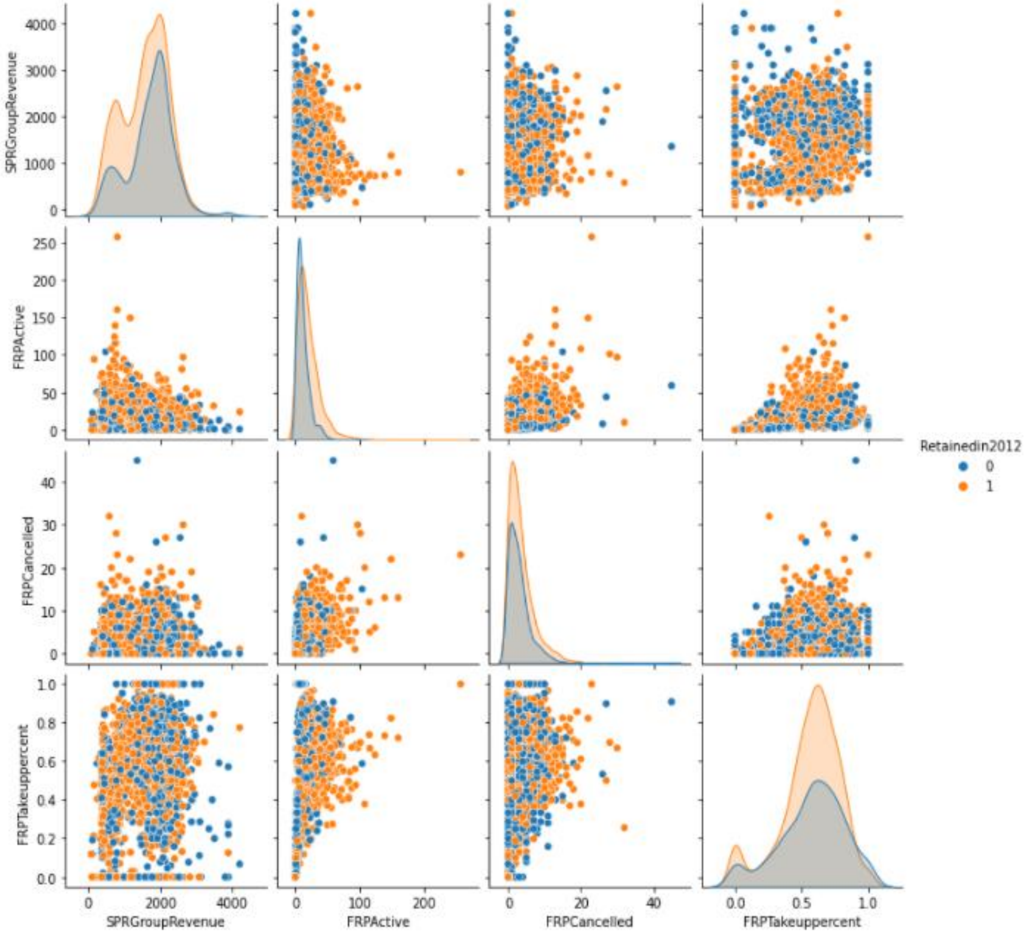
Different features through retention to check more deeply

- Total School Enrollment, Tuition, Deposit days before Departure, Days



Different features through retention to check more deeply

- SPR Group Revenue, FRP Active, FRP Cancelled, FRP Take up percent



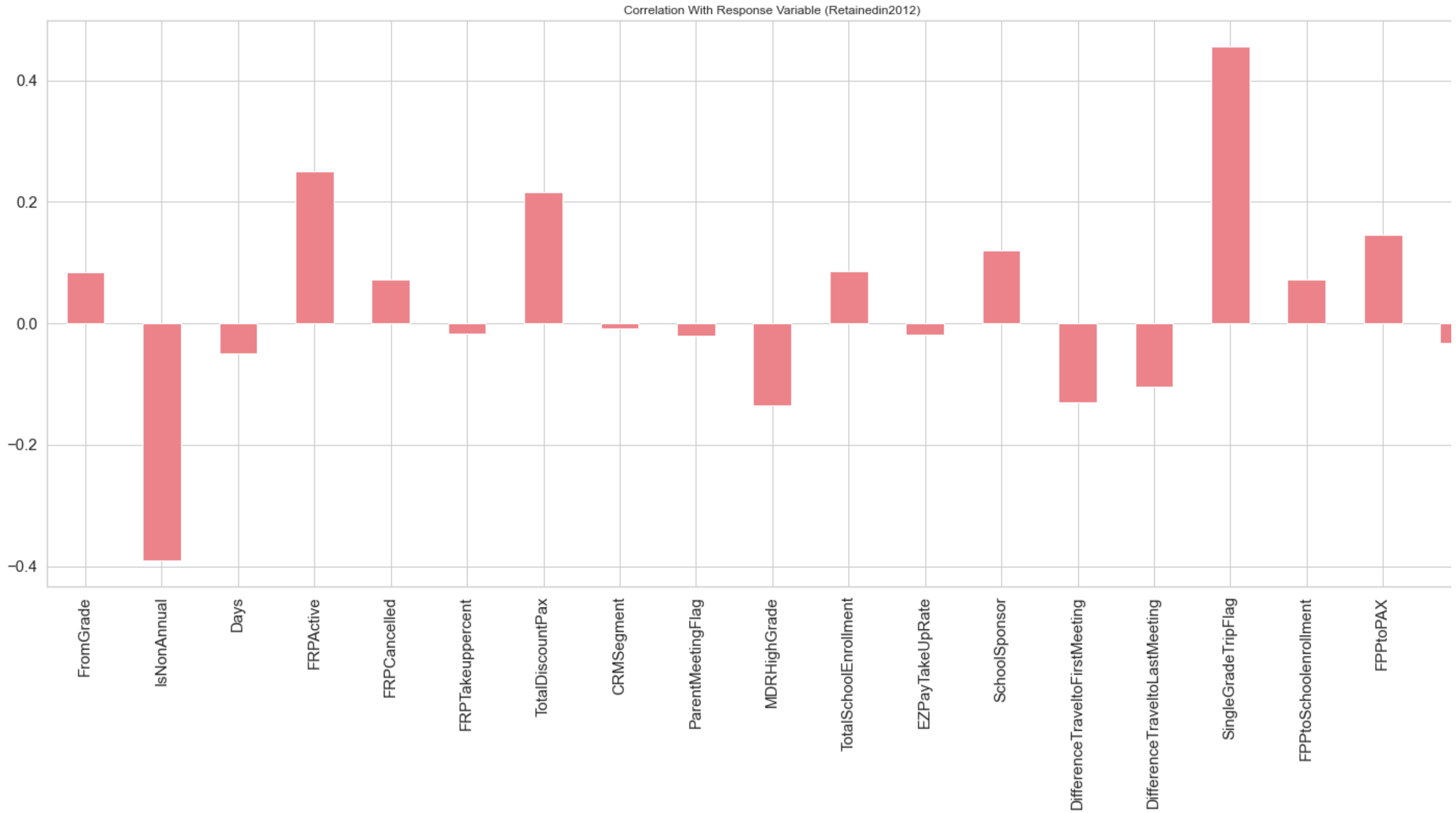
Most positive and negative correlated features with target value

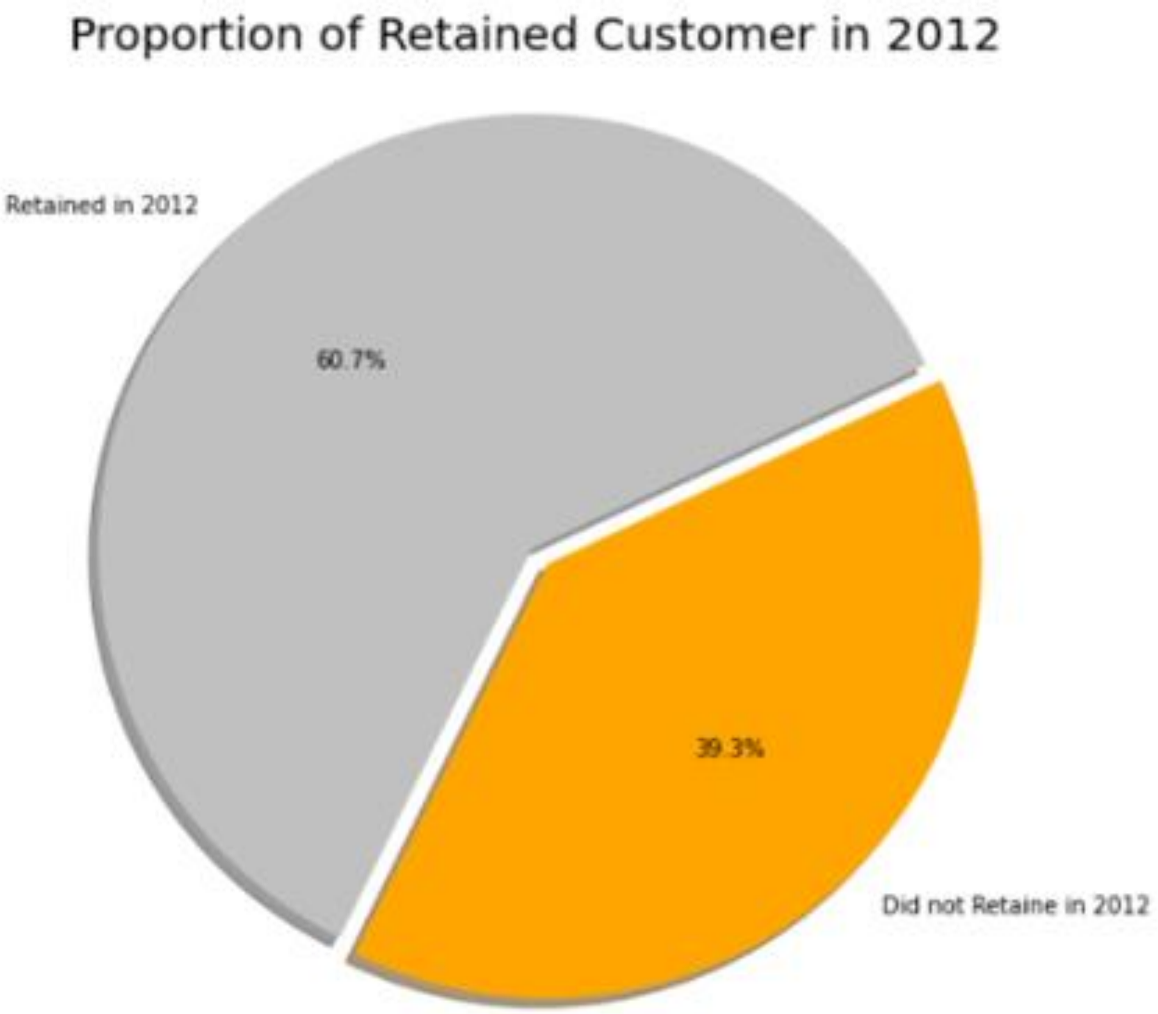
- **Interestingly, the Retention rate increases when**

1. full refund is active
2. If students were from the same grade
3. If the school is sponsoring the trip

- **In contrast features which seem to be negatively related to churn**

1. Days (the shortest the trip the most likely to retain)
2. In none Annual





Proportion of Retained Customer in 2012

```
   Retainedin2012
1            1451
0             938
```

Overall for the year 2012 it can be clearly seen that 60% of the customers retained in 2012.
While almost 40% of them churned.

Feature Engineering (Feature Selection)

After visualization analysis, an important step to take is dataset treatment. This helps conducting better feature.

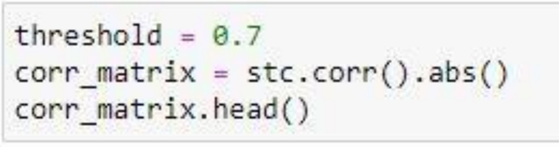
- *Remove Column with most Nan Values*
- *Remove Collinear Variables*
- *Removing Columns by Correlation With Response Variable*
- *Treat Missing Values*

Remove Column with most Nan Values

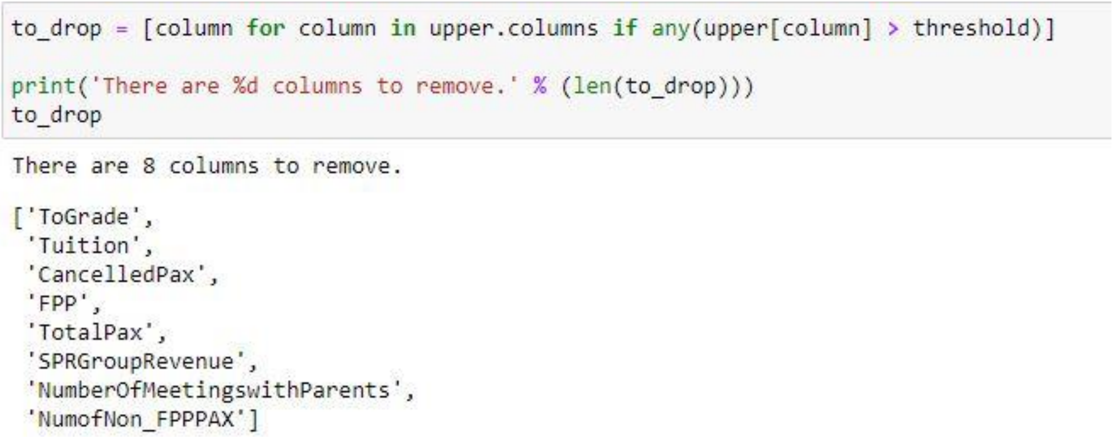
- *Special Pay and Early RPL with 1919 and 673 respectively have the most nan values.*

Remove Collinear Variables

- *Correlation matrix between each feature.*
- *Collinearity can reduce the performance of our models because it increases variance and the number of dimensions. We removed collinear variables as part of the data preparation process and eliminating variables that are above a certain threshold. For our project We select this threshold to be 70% .*



```
threshold = 0.7
corr_matrix = stc.corr().abs()
corr_matrix.head()
```



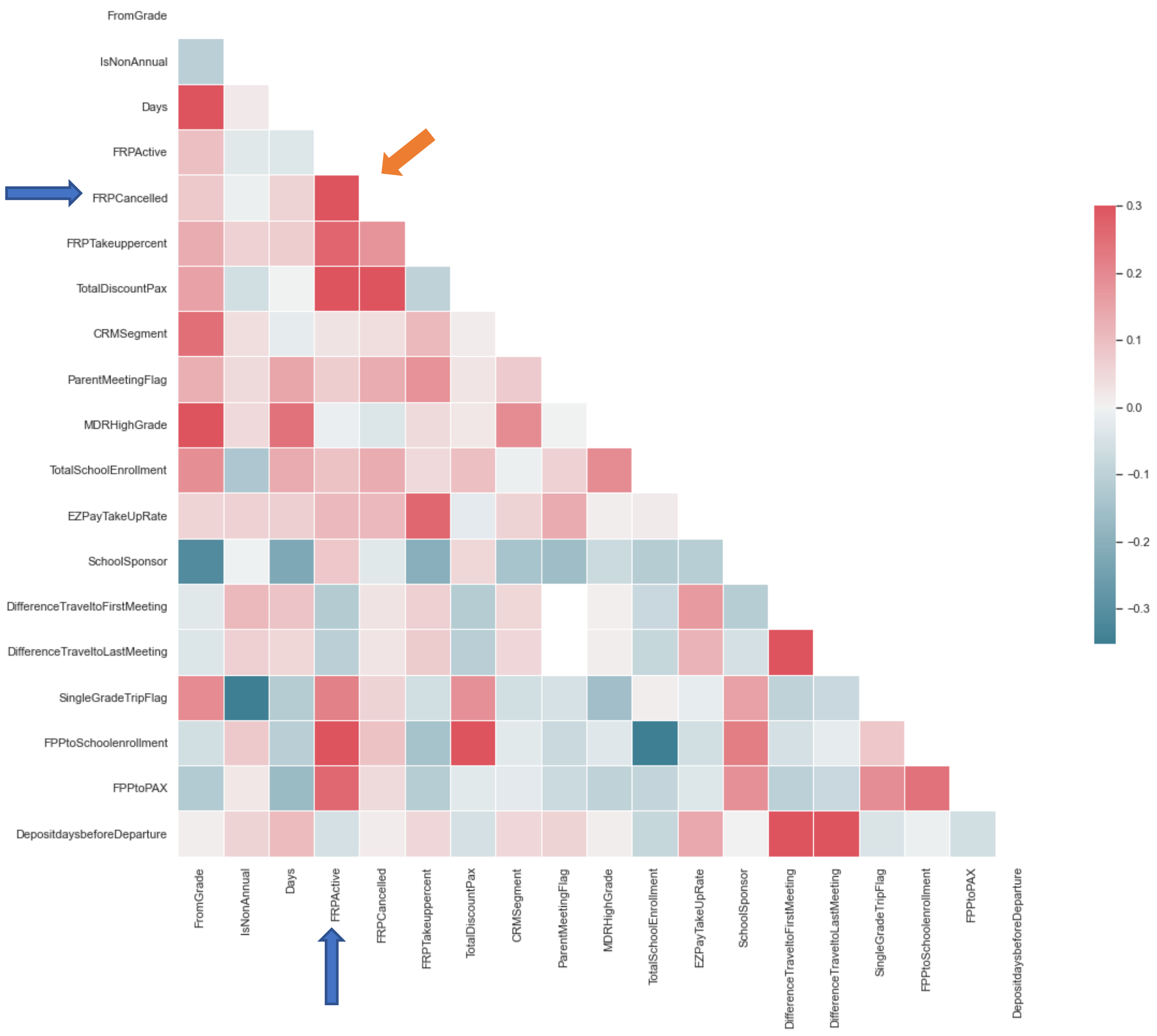
```
to_drop = [column for column in upper.columns if any(upper[column] > threshold)]

print('There are %d columns to remove.' % (len(to_drop)))
to_drop
```

```
There are 8 columns to remove.

['ToGrade',
 'Tuition',
 'CancelledPax',
 'FPP',
 'TotalPax',
 'SPRGroupRevenue',
 'NumberOfMeetingswithParents',
 'NumofNon_FPPPAX']
```

Removing Columns by Correlation With Response Variable



- Correlation matrix helps us to discover the bivariate relationship between independent variables in a dataset
- By removing a feature that has the most correlation with other feature, we will not lose the value of that, because we will still have the other feature.

For example: FRPCancelled is highly correlated with FRPActive. By removing one of them, we still can keep its value in another value.

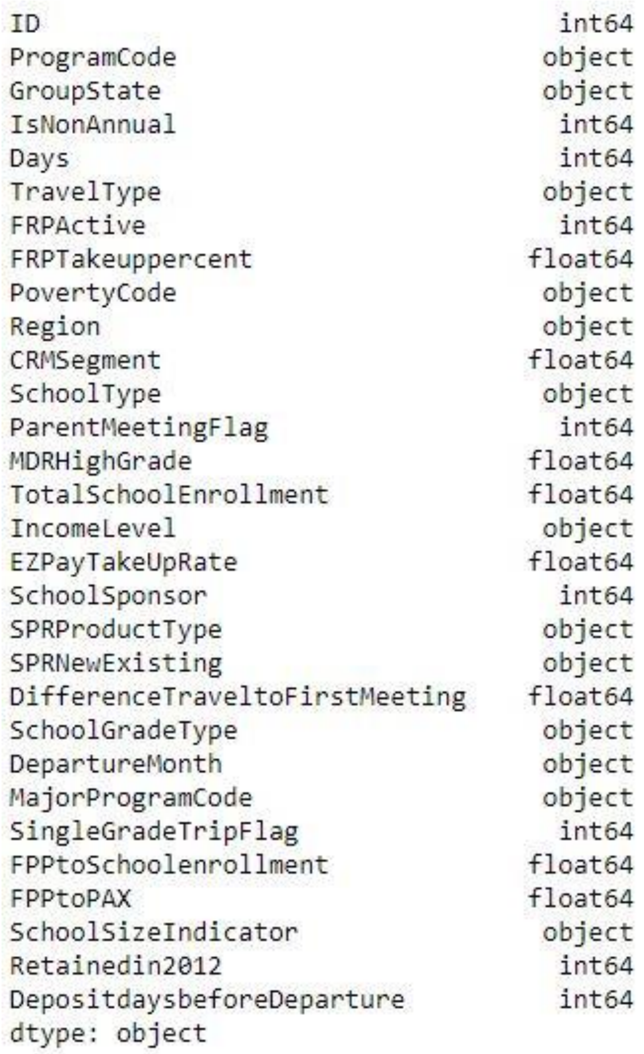
Moreover, highly correlated values are removed leading dropping four features

Treat Missing Values

- ✓ *For Numeric Values, Filling Nan values with Mean of its column*
- ✓ *For Categorical Values, Filling Nan values with Mode of its column*
- ✓ *Filling Nan in Date Columns with next available Non-Nan Value (to be used for further data visualization)*

- ❖ *Ignore Date Columns (only for applying classification method)*

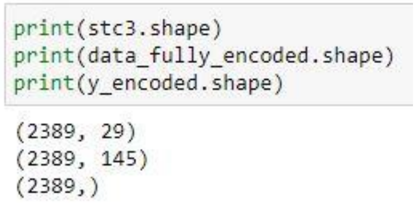
Final Selected Features



```
ID                                int64
ProgramCode                      object
GroupState                       object
IsNonAnnual                       int64
Days                              int64
TravelType                       object
FRPActive                         int64
FRPTakeuppercent                float64
PovertyCode                      object
Region                           object
CRMSegment                      float64
SchoolType                       object
ParentMeetingFlag                 int64
MDRHighGrade                    float64
TotalSchoolEnrollment           float64
IncomeLevel                      object
EZPayTakeUpRate                 float64
SchoolSponsor                     int64
SPRProductType                   object
SPRNewExisting                   object
DifferenceTraveltoFirstMeeting  float64
SchoolGradeType                  object
DepartureMonth                   object
MajorProgramCode                 object
SingleGradeTripFlag               int64
FPPtoSchoolenrollment           float64
FPPtoPAX                        float64
SchoolSizeIndicator              object
Retainedin2012                    int64
DepositdaysbeforeDeparture        int64
dtype: object
```

Preprocessing the Dataset

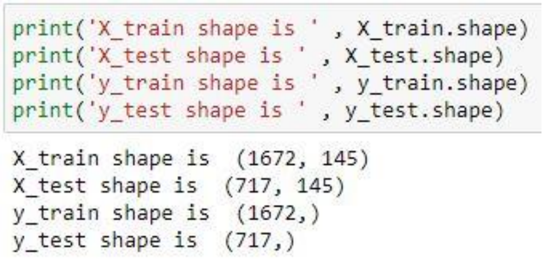
- *Encoding features (Ordinal and One_hot encoding)*



```
print(stc3.shape)
print(data_fully_encoded.shape)
print(y_encoded.shape)
```

```
(2389, 29)
(2389, 145)
(2389,)
```

- *Splitting encoded dataset into train and test set*

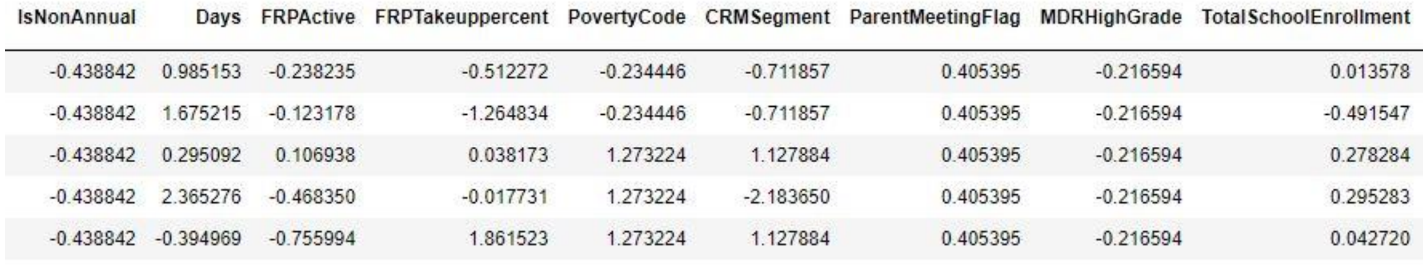


```
print('X_train shape is ' , X_train.shape)
print('X_test shape is ' , X_test.shape)
print('y_train shape is ' , y_train.shape)
print('y_test shape is ' , y_test.shape)
```

```
X_train shape is  (1672, 145)
X_test shape is  (717, 145)
y_train shape is  (1672,)
y_test shape is  (717,)
```

- *Conduct Feature Scaling*

✓ It’s quite important to normalize the variables before conducting any machine learning (classification) algorithms so that all the training and test variables are scaled within a range



IsNonAnnual	Days	FRPActive	FRPTakeuppercent	PovertyCode	CRMSegment	ParentMeetingFlag	MDRHighGrade	TotalSchoolEnrollment
-0.438842	0.985153	-0.238235	-0.512272	-0.234446	-0.711857	0.405395	-0.216594	0.013578
-0.438842	1.675215	-0.123178	-1.264834	-0.234446	-0.711857	0.405395	-0.216594	-0.491547
-0.438842	0.295092	0.106938	0.038173	1.273224	1.127884	0.405395	-0.216594	0.278284
-0.438842	2.365276	-0.468350	-0.017731	1.273224	-2.183650	0.405395	-0.216594	0.295283
-0.438842	-0.394969	-0.755994	1.861523	1.273224	1.127884	0.405395	-0.216594	0.042720

All values in each column are Balanced

Applying Classification Algorithm

- *What is done so far was preparing the dataset so we can apply classification methods on them. The rest of the presentation illustrates how 5 different classification method is applied to the dataset, measure their accuracy with* ***confusion matrix*** *help which is followed by comparing them and deciding which one to choose. Five Classification algorithms are listed below:*

- *Logistic Regression*
- *Decision Tree*
- *Random Forest*
- *K Nearest Neighbor (KNN)*
- *Support Vector Machine (SVM)*
- XGBoost

- *Accuracy of mode baseline model =24.83*

Confusion Matrix

Retained = Yes or 1
Do not Retain = No or 0

A confusion matrix is a table that is often used to **describe the performance of a classification model**

Accuracy: Overall, how often is the classifier correct?

True Positive Rate: When it's Actually yes, how often does it predict yes?

False Positive Rate (Type 1 Error): When it's Actually no, how often does it predict yes?

True Negative Rate: When it's Actually no, how often does it predict no?

False Negative Rate(Type 2 Error): When it's Actually yes, how often does it predict no?

Precision: When it predicts yes, how often is it correct?

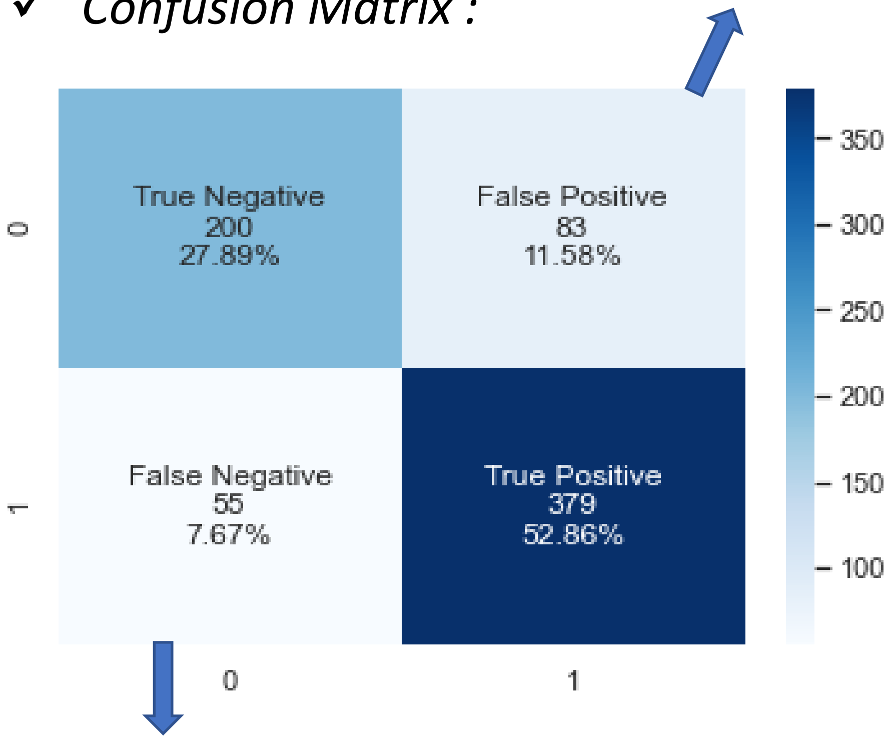
F Score: Weighted average of the true positive rate (recall) and precision

Also, it is used to check where our classification model fails

Logistic Regression

- ✓ *Training Set Accuracy = 81.51 %*
- ✓ *Test Set Accuracy = 80.75 %*
- ✓ *Confusion Matrix :*

Type 1 error



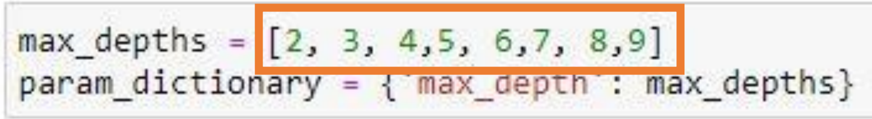
Type 2 error

	Precision	Recall	F1-Score	Support
Yes	82%	87%	85%	434
No	78%	71%	74%	283

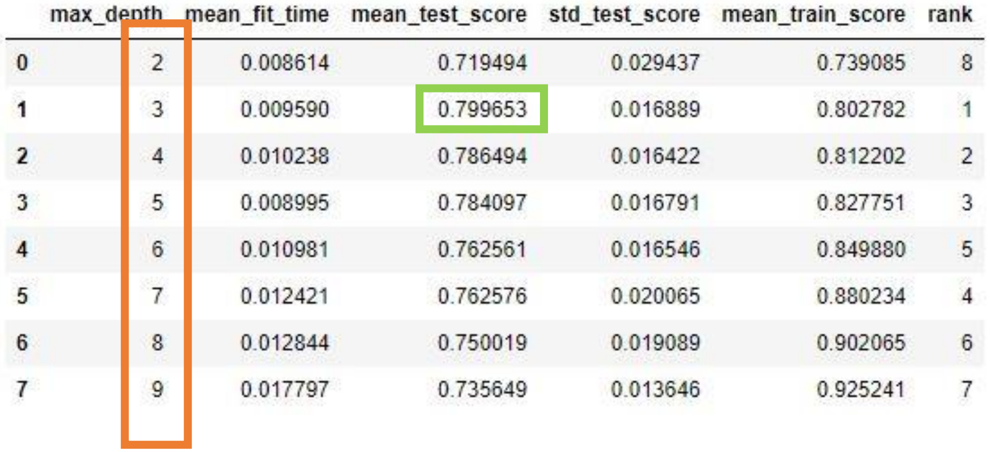
- ✓ *Accuracy of Logistic Regression = 80.75 %*

Decision Tree

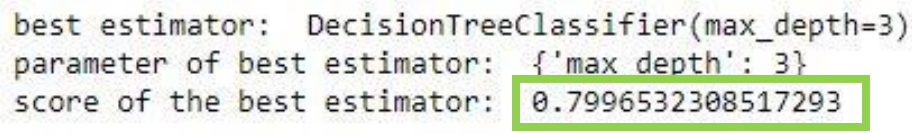
✓ *Max Depth Identifier*



```
max_depths = [2, 3, 4,5, 6,7, 8,9]
param_dictionary = {'max_depth': max_depths}
```



	max_depth	mean_fit_time	mean_test_score	std_test_score	mean_train_score	rank
0	2	0.008614	0.719494	0.029437	0.739085	8
1	3	0.009590	0.799653	0.016889	0.802782	1
2	4	0.010238	0.786494	0.016422	0.812202	2
3	5	0.008995	0.784097	0.016791	0.827751	3
4	6	0.010981	0.762561	0.016546	0.849880	5
5	7	0.012421	0.762576	0.020065	0.880234	4
6	8	0.012844	0.750019	0.019089	0.902065	6
7	9	0.017797	0.735649	0.013646	0.925241	7

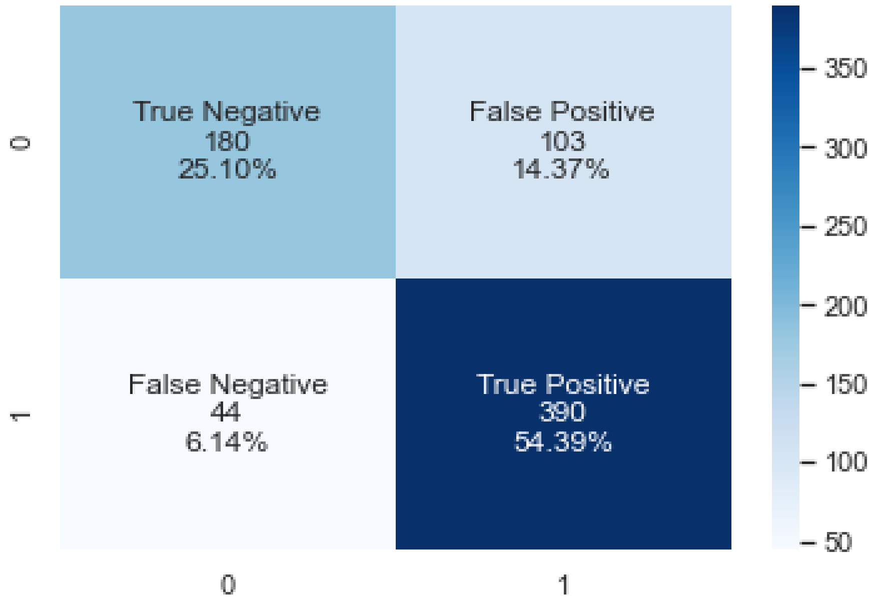


```
best estimator:  DecisionTreeClassifier(max_depth=3)
parameter of best estimator:  {'max_depth': 3}
score of the best estimator:  0.7996532308517293
```

Depth equal to 3 gives us the most accuracy

Decision Tree

- ✓ *Training Set Accuracy = 100 %*
- ✓ *Test Set Accuracy = 73.08 %*
- ✓ *Confusion Matrix :*

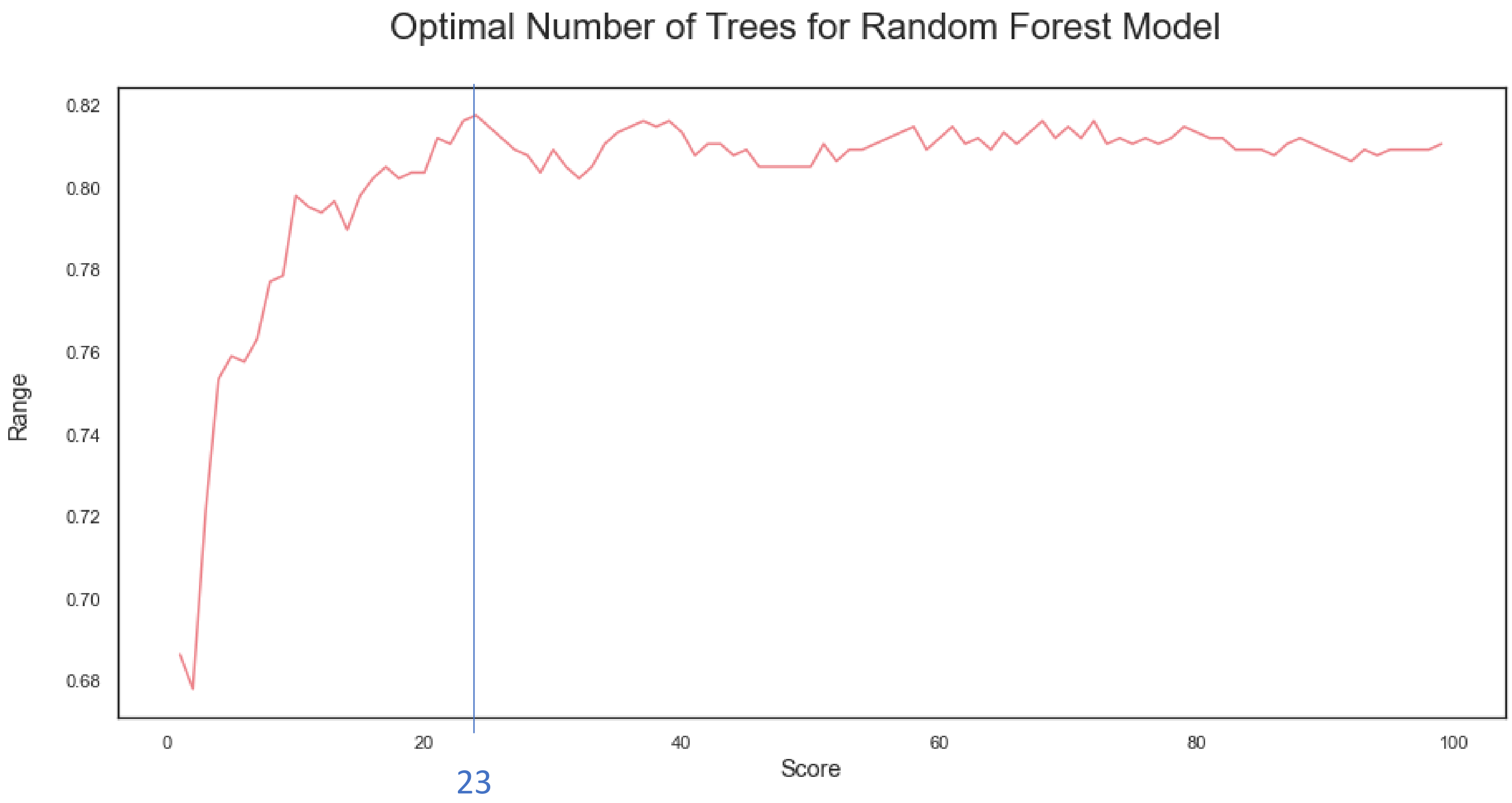


	Precision	Recall	F1-Score	Support
Yes	80%	90%	84%	434
No	80%	64%	71%	283

- ✓ *Accuracy of Decision Tree= 79.49 %*

Random Forest

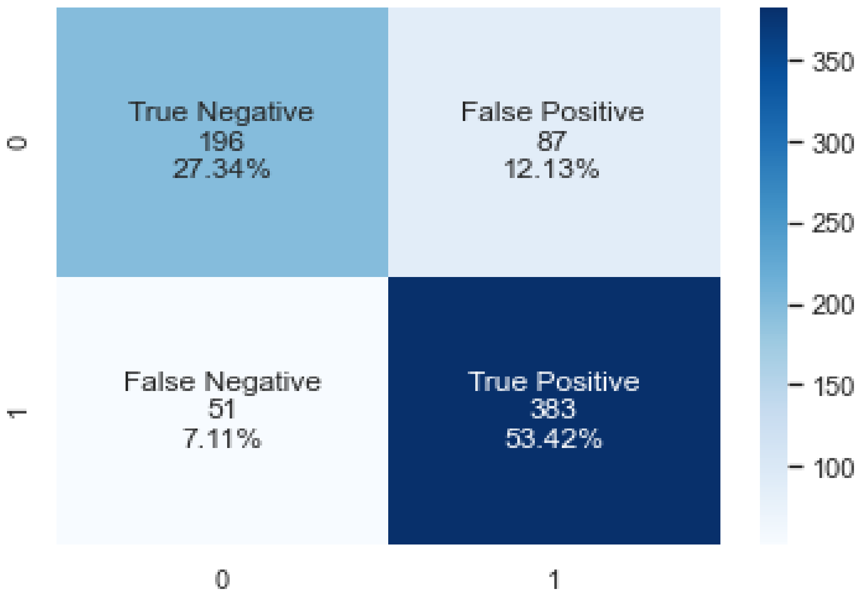
- ✓ *Optimal Number of Trees*



As we could see from the iterations above, the random forest model would attain the highest accuracy score when its n_estimators = 23

Random Forest

- ✓ *Training Set Accuracy = 100 %*
- ✓ *Test Set Accuracy = 81.31 %*
- ✓ *Confusion Matrix :*

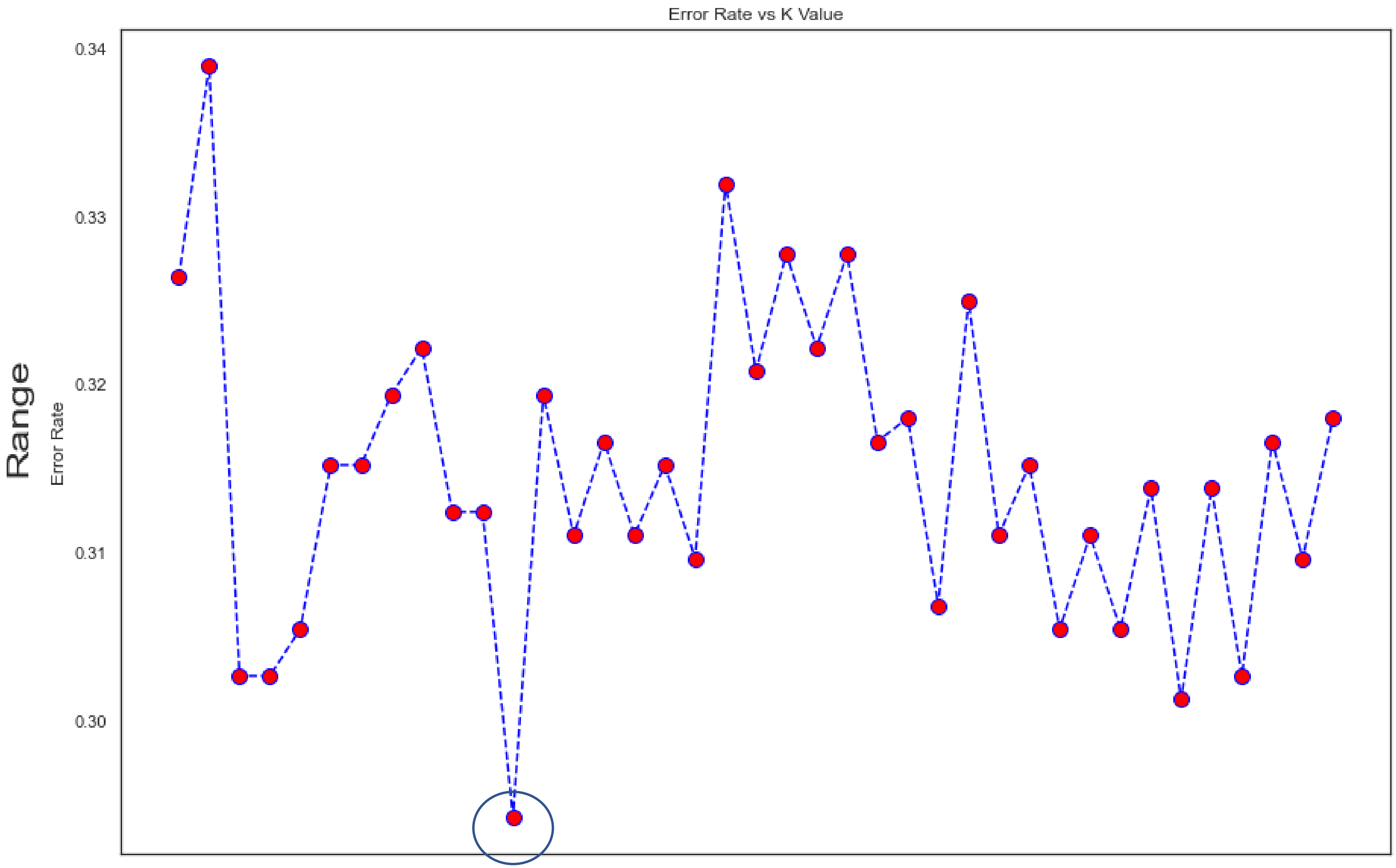


	Precision	Recall	F1-Score	Support
Yes	81%	88%	85%	434
No	79%	69%	74%	283

- ✓ *Accuracy of Random Forest= 80.75 %*

K-Nearest Neighbor

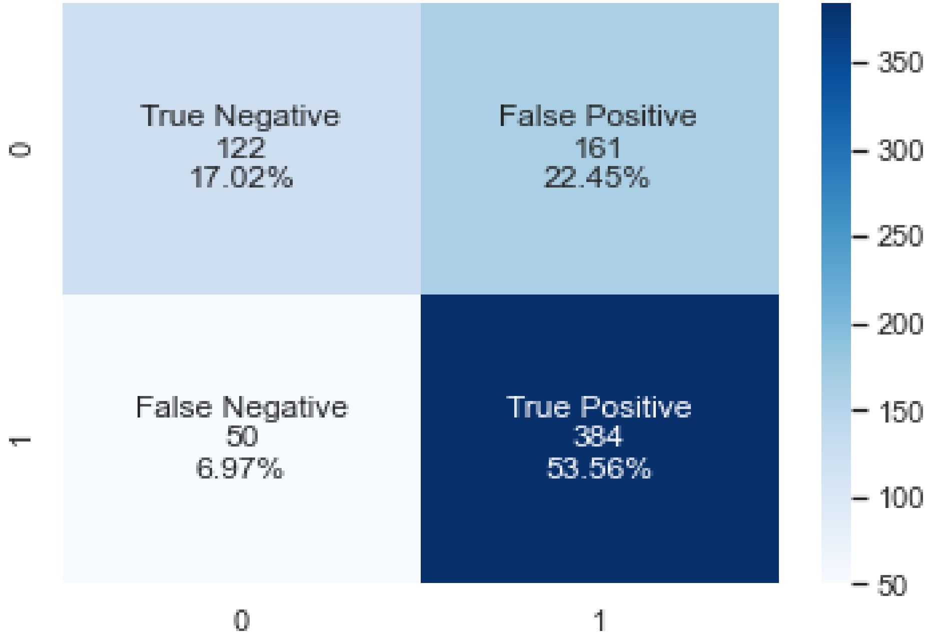
✓ Identify the optimal number of K neighbors for KNN Model



Optimal Number of K = 12

K-Nearest Neighbor

- ✓ *Training Set Accuracy = 77.57 %*
- ✓ *Test Set Accuracy = 69.45 %*
- ✓ *Confusion Matrix :*

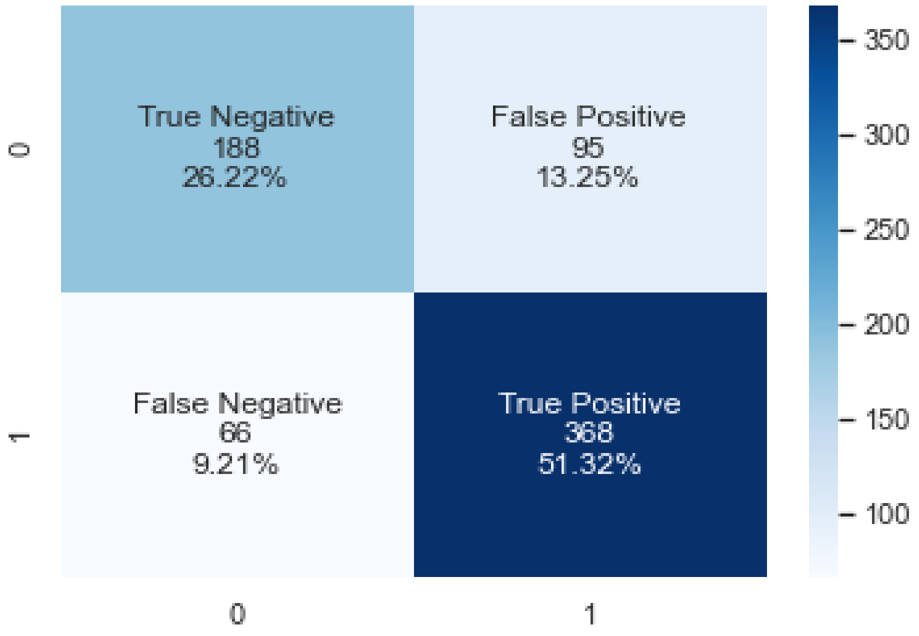


	Precision	Recall	F1-Score	Support
Yes	70%	88%	78%	434
No	71%	43%	54%	283

- ✓ *Accuracy of K-Nearest Neigbhor= 70.57 %*

Support Vector Machine

- ✓ *Training Set Accuracy = 86.24 %*
- ✓ *Test Set Accuracy = 77.82 %*
- ✓ *Confusion Matrix :*

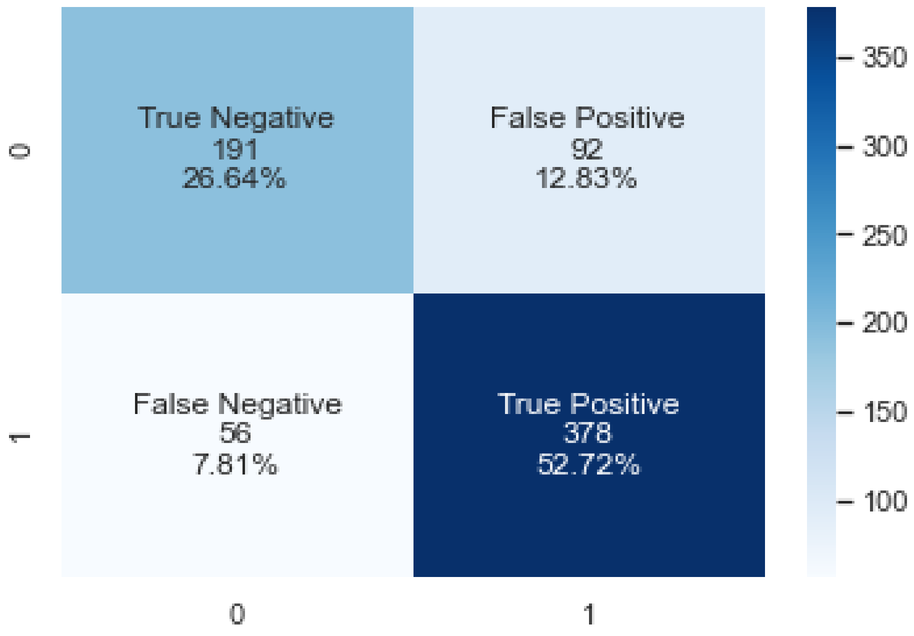


	Precision	Recall	F1-Score	Support
Yes	79%	85%	82%	434
No	74%	66%	70%	283

- ✓ *Accuracy of Support Vector Machine= 77.54 %*

XGBoost

- ✓ *Training Set Accuracy = 92 %*
- ✓ *Test Set Accuracy = 79 %*
- ✓ *Confusion Matrix :*

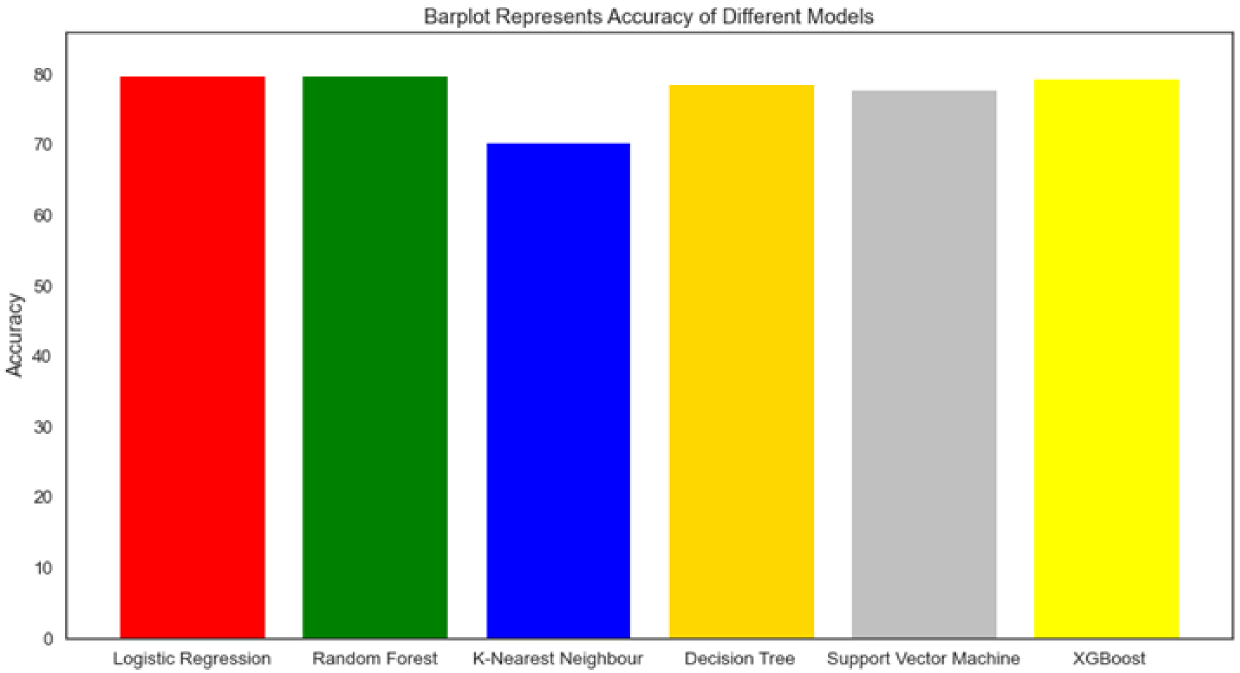


	Precision	Recall	F1-Score	Support
Yes	80%	87%	84%	434
No	77%	67%	72%	283

- ✓ *Accuracy of XGBoost = 79.35 %*

Choosing the Best Classification Algorithm

	Model	Accuracy
1	Logistic Regression	79.474198
2	Random Forest	80.055788
3	KNN	71.827057
4	Decision Tree	79.497908
5	SVM	77.872553
6	XGB	80.753138



Overall, every classification algorithms has relatively the same accuracy. XGBoost and random forest are the best model for the given dataset because they have the best combination of precision, recall, and F2 scores, resulting in the most correct positive predictions with the fewest false negatives. Performance valuating of the chosen algorithms is the next step.

Choosing the Best Classification Algorithm

	Model	Accuracy
1	Logistic Regression	79.474198
2	Random Forest	80.055788
3	KNN	71.827057
4	Decision Tree	79.497908
5	SVM	77.872553
6	XGB	80.753138

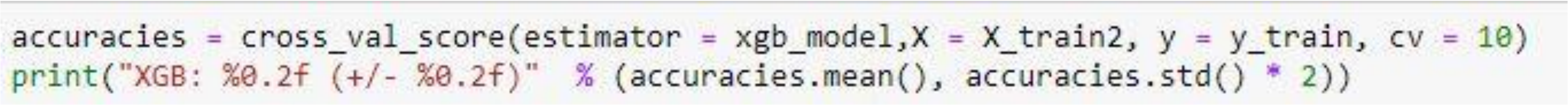
	Model	Accuracy
1	Logistic Regression	79.637378
2	Random Forest	81.450488
3	KNN	70.013947
4	Decision Tree	79.497908
5	SVM	77.872553
6	XGB	79.765438

Overall, every classification algorithms has relatively the same accuracy. XGBoost and random forest are the best model for the given dataset because they have the best combination of precision, recall, and F2 scores, resulting in the most correct positive predictions with the fewest false negatives. Performance valuating of the chosen algorithms is the next step.

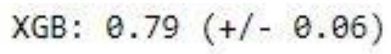
Model Evaluation

K- fold Cross-Validation

Model evaluation using '**K- fold Cross-Validation**' technique that helps us to fix the variance. In order to fix the variance problem, k-fold cross-validation basically split the training set into 10 folds and train the model on 9 folds (9 subsets of the training dataset).



```
accuracies = cross_val_score(estimator = xgb_model,X = X_train2, y = y_train, cv = 10)
print("XGB: %0.2f (+/- %0.2f)"  % (accuracies.mean(), accuracies.std() * 2))
```



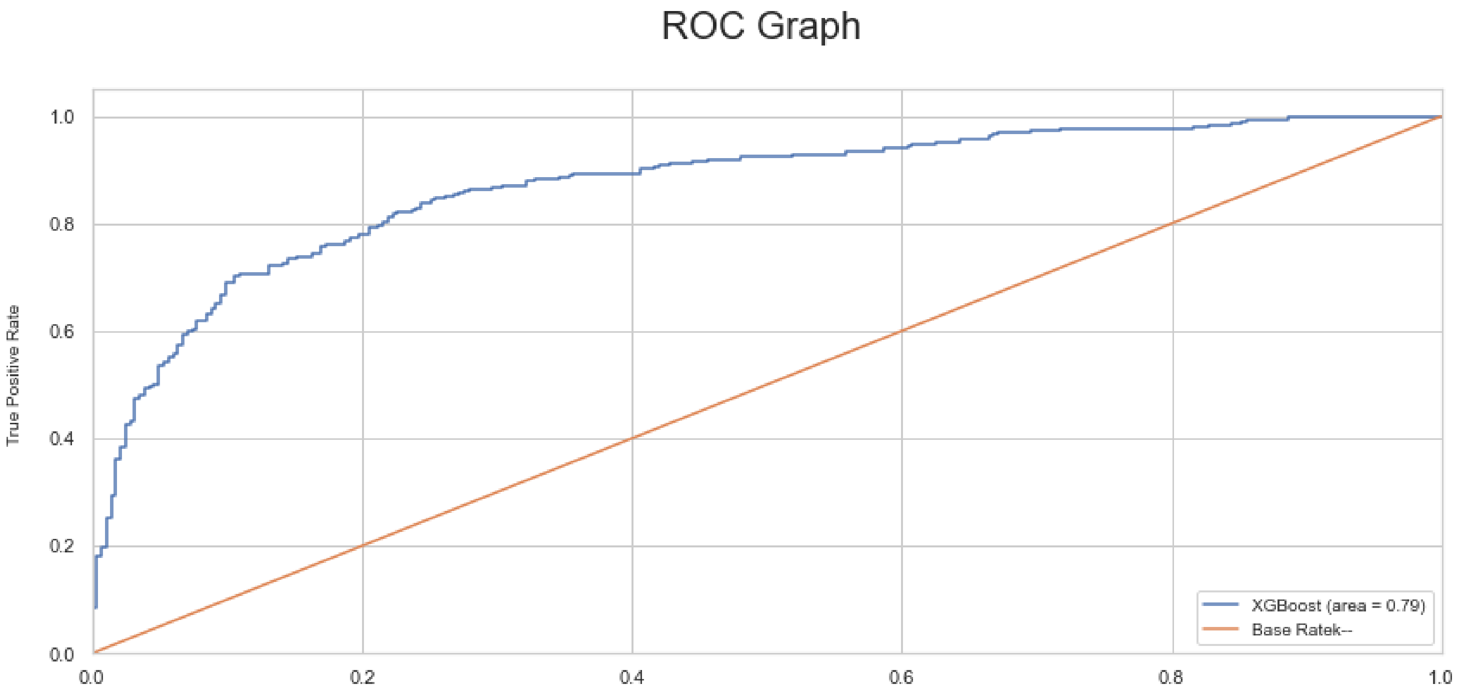
```
XGB: 0.79 (+/- 0.06)
```

Therefore, our k-fold Cross Validation results indicate that we would have an accuracy anywhere between 73% to 85% while running this model on any test set.

Model Evaluation

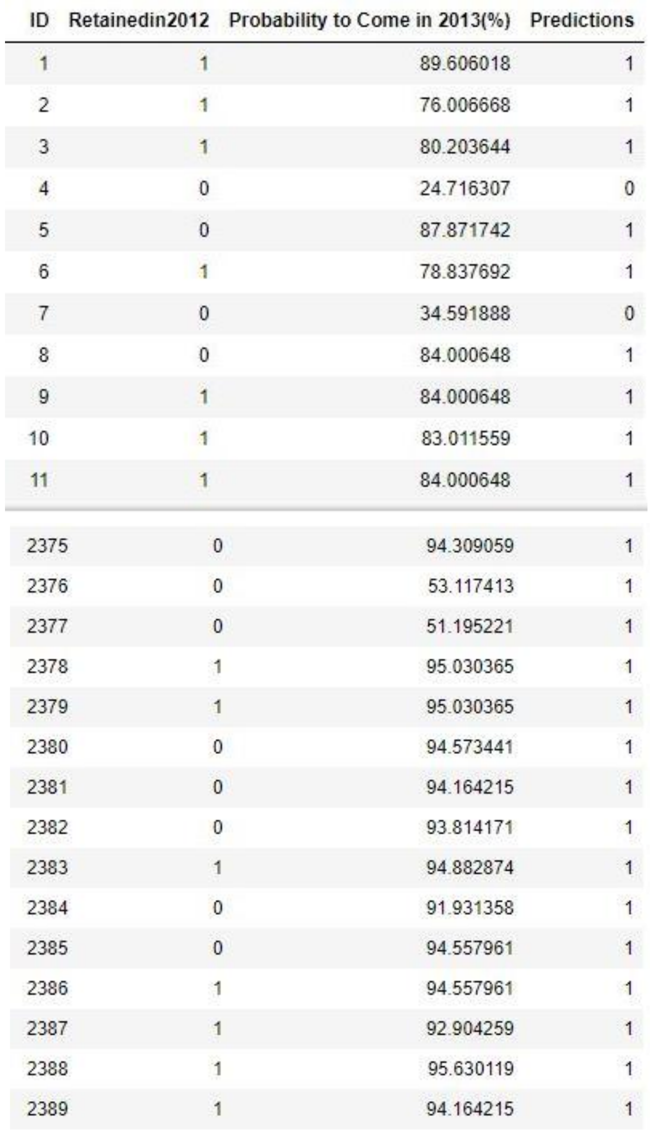
ROC Graph

It’s good to re-evaluate the model using ROC Graph. ROC Graph shows us the capability of a model to distinguish between the classes based on the AUC Mean score. The orange line represents the ROC curve of a random classifier while a good classifier tries to remain as far away from that line as possible. As shown in the graph below, the fine-tuned XGBoost model showcased a higher AUC score.



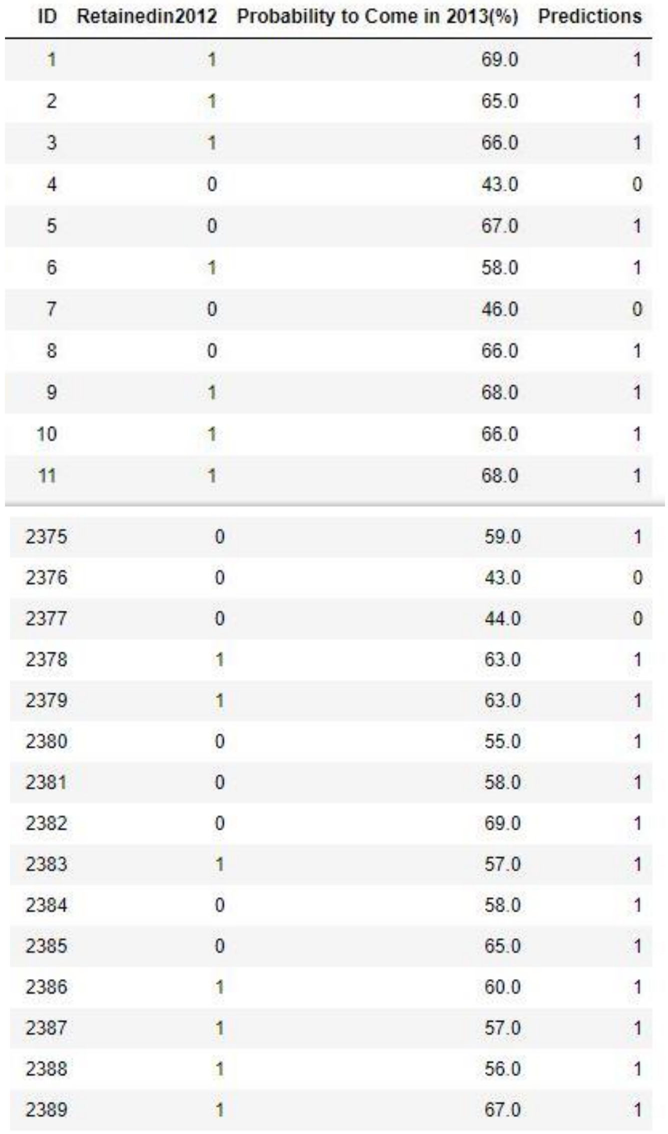
Forecasting the Following Year

XGB

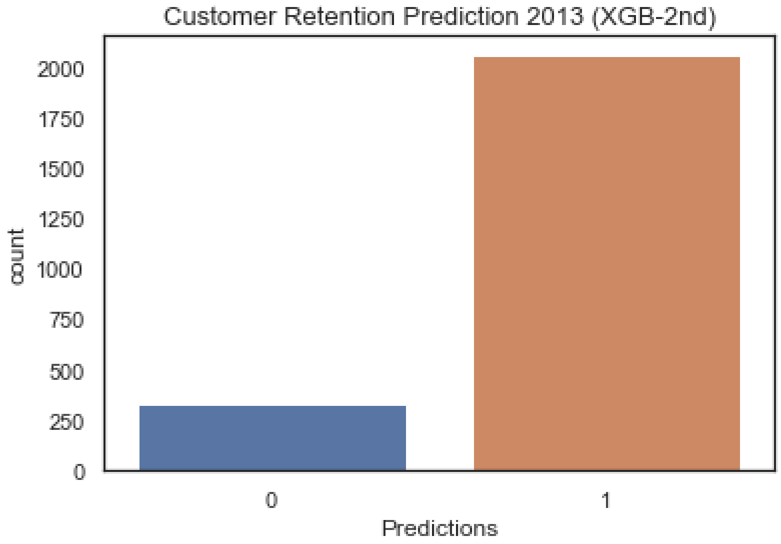


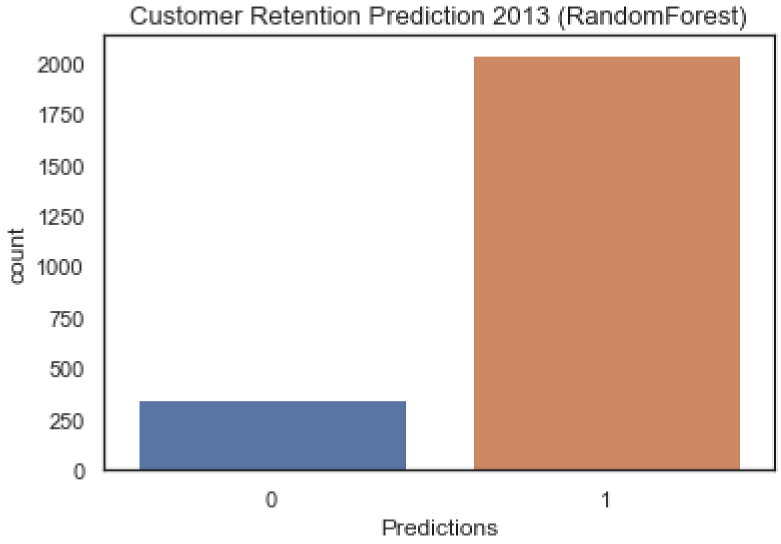
ID	Retainedin2012	Probability to Come in 2013(%)	Predictions
1	1	89.606018	1
2	1	76.006668	1
3	1	80.203644	1
4	0	24.716307	0
5	0	87.871742	1
6	1	78.837692	1
7	0	34.591888	0
8	0	84.000648	1
9	1	84.000648	1
10	1	83.011559	1
11	1	84.000648	1
2375	0	94.309059	1
2376	0	53.117413	1
2377	0	51.195221	1
2378	1	95.030365	1
2379	1	95.030365	1
2380	0	94.573441	1
2381	0	94.164215	1
2382	0	93.814171	1
2383	1	94.882874	1
2384	0	91.931358	1
2385	0	94.557961	1
2386	1	94.557961	1
2387	1	92.904259	1
2388	1	95.630119	1
2389	1	94.164215	1

Random Forest



ID	Retainedin2012	Probability to Come in 2013(%)	Predictions
1	1	69.0	1
2	1	65.0	1
3	1	66.0	1
4	0	43.0	0
5	0	67.0	1
6	1	58.0	1
7	0	46.0	0
8	0	66.0	1
9	1	68.0	1
10	1	66.0	1
11	1	68.0	1
2375	0	59.0	1
2376	0	43.0	0
2377	0	44.0	0
2378	1	63.0	1
2379	1	63.0	1
2380	0	55.0	1
2381	0	58.0	1
2382	0	69.0	1
2383	1	57.0	1
2384	0	58.0	1
2385	0	65.0	1
2386	1	60.0	1
2387	1	57.0	1
2388	1	56.0	1
2389	1	67.0	1





Conclusion

- For the next year, it is concluded that the number of retained customer will increase.
- Feature Selection as an initial step for classification method is of great importance. By selecting different combination of features together, classification algorithms got different accuracy scores which lead to getting not equal probabilities of retention for each customer.
- Overall, two different group of selected features resulted in the same prediction but with different accuracies.

- **In General**

- According to low sponsoring rate of schools, company need to consider if they want to accept sponsors or if they can investigate the money spent to find school sponsors in different area such as absorbing new customers.
- Although flying is more popular than other vehicles, but their share remains between both new and existing. For this reason, it is recommended that the distance feature be considered in future records as well, as it is possible that customers only want to use the aircraft for long distances and prefer buses or trains for shorter distances.
- Since region ’other’ has the highest participant, company need to reconsider data collection according to this feature.
- Duo to company’s strategy, we suggest to have periodic short-term activities in order to enables us add more measurable features and better investigation
- the most crucial factors for customers retention are FRP Active,TotalDiscountPax, Single Grade Trip. This results are gained through correlation analysis

- **for Marketing department according to data**

- Focusing more on travel type A and B while marketing can consider if they are able to change the concept of travel type for instance, T and N (because of the low rate of either participation or retention
- There is a need to absorb new customers according to their nature as well as their field of activity which leads to loyalty.
- Poverty code (B,C,A) showed the most interest to travel, so marketing team can focus on working with them with the same behavior, while as per poverty code (D,E,O) company can provide them with the promotions to check either they change their behavior, or this is the nature of these specific group of customer.
- Public Schools had the most retention rate while it also had the least amount of parent meetings, so there is a need of more parent meeting to see if an increase of this rate can be seen.
- Most number of new costumers are absorbed when the destination was Eastcoast, which can guide the marketing to find out how they can attract more new customers to the system.
- School type elementary- high had the highest level of participation in the company’s product, which means, company needs to work more on attracting other school types or limit their customer just to this type of schools and take the best out of them.

References

- Customer churn: A study of factors affecting customer churn using machine learning https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=1286&context=creativecomponents
- Customer churn www.kaggale.com
- www.youtube.com
- https://www.lorensworld.com/business/put-your-skills-to-the-test-customer-service-101/