Gro

Retention Modeling at STc

Harvard Case Study Project

Phase 2 Submission

Group 9

# Data Description and Quality

## Data Description

Based on our exploration of the data types, we have the following constitution in the two tables:

**Exhibit 2:**

|  |  |  |
| --- | --- | --- |
| **Data type** | **Columns** | **Count** |
| Categorical | ID, Program Code, Group State, Is Non Annual , Travel Type, Special Pay, Poverty Code, Region, CRM Segment, School Type, Parent Meeting Flag, Income Level, School Sponsor, SPR Product Type, SPR New Existing, SchoolGradeTypeLow, SchoolGradeTypeHigh, SchoolGradeType, DepartureMonth, GroupGradeTypeLow, GroupGradeTypeHigh, GroupGradeType, MajorProgramCode, SingleGradeTripFlag, SchoolSizeIndicator, Retained in 2012 | 26 |
| Numerical | From Grade, To Grade, Days, Tuition, FRP Active, FRP Cancelled, FRP Take up percent , Cancelled Pax, Total Discount Pax, MDR Low Grade, MDR High Grade, Total School Enrollment, EZ Pay Take Up Rate, FPP, Total Pax, SPR Group Revenue, NumberOfMeetingswithParents, DifferenceTraveltoFirstMeeting, DifferenceTraveltoLastMeeting, FPP to School enrollment, FPP to PAX, Num of Non\_FPP PAX | 22 |
| DateTime | Departure Date, Return Date, Deposit Date, Early RPL, Latest RPL, Initial System Date, FirstMeeting, LastMeeting | 8 |

**Exhibit 1:**

|  |  |  |
| --- | --- | --- |
| **Data type** | **Columns** | **Count** |
| Numerical | NPS Score - 2011, NPS Score - 2011, NPS Score - 2009, NPS Score - 2009 | 4 |
| DateTime | >= 3 FPP Date, >= 10 FPP Date, >=20 FPP Date, >=35 FPP Date | 4 |

## Data Quality

After having gone through the data, we have the following comments:

1. We have conducted some tests on the data to check its quality. We looked for the following in our tests:
   1. Range within expected values. (E.g.: Proportions between 0 and 1)
   2. Proportion of missing data in a variable.
2. We have no way of confirming or denying the efficacy of the process through which this data was generated. We merely assumed that the data is of high quality, and test our assumption against the magnitude of its predictive power.

# Technologies used

For this project, we relied on 2 tools:

1. **R** – Our team was familiar with R and we used this for every step in Section 3 of this report. Time was a critical element and it was easier for us to stick to R for the intricate and complex steps needed in pre-processing.
2. **SAS JMP** – We used SAS JMP for the actual modelling stage because our team has used this tool before as part of Project 1.

# Data Pre-processing

We will now describe the pre-processing steps we took, in the order we took them.

## Understanding Missing Values

This step involved glancing at the data and looking at the different values contained in the categorical and numeric columns. When we did so,

**we found that many of the columns had the string “NA” in them.**

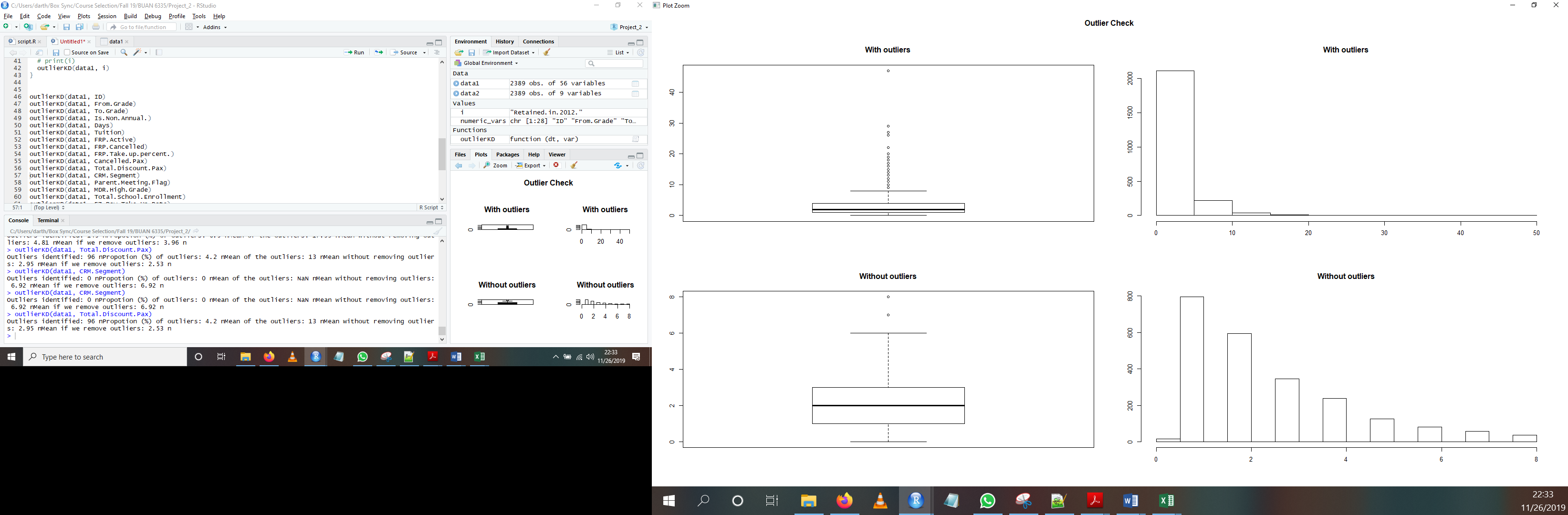
We were able to isolate the columns in which the value “NA” was legitimate and did mean something versus those where NA was to be replaced with a missing value, later-on treated in the imputation step.

The next table highlights our findings. The cells highlighted green are the ones in which the “NA” values have meaning. In the remaining, the “NA” values need to be removed for processing later.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Exhibit 2** | | |  | **Exhibit 1** | | |
| **Column** | **"NA" Values** | **Missing Values** |  | **Column** | **"NA" Values** | **Missing Values** |
| Special.Pay | 1917 | 1919 |  | >= 35 FPP Date | 0 | 1618 |
| Early.RPL | 673 | 673 |  | NPS 2008 | 0 | 1412 |
| Poverty.Code | 0 | 599 |  | NPS 2010 | 0 | 1229 |
| FirstMeeting | 337 | 337 |  | NPS 2009 | 0 | 1225 |
| DifferenceTraveltoFirstMeeting | 337 | 337 |  | >= 20 FPP Date | 0 | 1024 |
| LastMeeting | 337 | 337 |  | NPS 2011 | 0 | 577 |
| DifferenceTraveltoLastMeeting | 337 | 337 |  | >= 10 FPP Date | 0 | 409 |
| To.Grade | 150 | 150 |  | >= 3 FPP Date | 0 | 9 |
| From.Grade | 127 | 127 |  | ID | 0 | 0 |
| FPP.to.School.enrollment | 91 | 91 |  |  |  |  |
| Total.School.Enrollment | 0 | 91 |  |  |  |  |
| SchoolSizeIndicator | 0 | 91 |  |  |  |  |
| MDR.High.Grade | 68 | 68 |  |  |  |  |
| MDR.Low.Grade | 0 | 68 |  |  |  |  |
| Income.Level | 0 | 62 |  |  |  |  |
| Latest.RPL | 19 | 19 |  |  |  |  |
| Initial.System.Date | 8 | 8 |  |  |  |  |
| CRM.Segment | 4 | 4 |  |  |  |  |
| ID | 0 | 0 |  |  |  |  |
| Group.State | 0 | 0 |  |  |  |  |
| Days | 0 | 0 |  |  |  |  |
| Departure.Date | 0 | 0 |  |  |  |  |
| Deposit.Date | 0 | 0 |  |  |  |  |
| Tuition | 0 | 0 |  |  |  |  |
| FRP.Cancelled | 0 | 0 |  |  |  |  |
| Cancelled.Pax | 0 | 0 |  |  |  |  |
| Region | 0 | 0 |  |  |  |  |
| School.Type | 0 | 0 |  |  |  |  |
| EZ.Pay.Take.Up.Rate | 0 | 0 |  |  |  |  |
| SPR.Product.Type | 0 | 0 |  |  |  |  |
| FPP | 0 | 0 |  |  |  |  |
| SPR.Group.Revenue | 0 | 0 |  |  |  |  |
| SchoolGradeTypeLow | 0 | 0 |  |  |  |  |
| SchoolGradeType | 0 | 0 |  |  |  |  |
| GroupGradeTypeLow | 0 | 0 |  |  |  |  |
| GroupGradeType | 0 | 0 |  |  |  |  |
| SingleGradeTripFlag | 0 | 0 |  |  |  |  |
| FPP.to.PAX | 0 | 0 |  |  |  |  |
| Program.Code | 0 | 0 |  |  |  |  |
| Is.Non.Annual. | 0 | 0 |  |  |  |  |
| Travel.Type | 0 | 0 |  |  |  |  |
| Return.Date | 0 | 0 |  |  |  |  |
| FRP.Active | 0 | 0 |  |  |  |  |
| FRP.Take.up.percent. | 0 | 0 |  |  |  |  |
| Total.Discount.Pax | 0 | 0 |  |  |  |  |
| Parent.Meeting.Flag | 0 | 0 |  |  |  |  |
| School.Sponsor | 0 | 0 |  |  |  |  |
| SPR.New.Existing | 0 | 0 |  |  |  |  |
| Total.Pax | 0 | 0 |  |  |  |  |
| NumberOfMeetingswithParents | 0 | 0 |  |  |  |  |
| SchoolGradeTypeHigh | 0 | 0 |  |  |  |  |
| DepartureMonth | 0 | 0 |  |  |  |  |
| GroupGradeTypeHigh | 0 | 0 |  |  |  |  |
| MajorProgramCode | 0 | 0 |  |  |  |  |
| Num.of.Non\_FPP.PAX | 0 | 0 |  |  |  |  |
| Retained.in.2012. | 0 | 0 |  |  |  |  |

## Studying Outliers

Studying outliers was the next big step we took in our analysis.



We studied the distribution of the 22 numerical and used intuition to assess the following:

* Are the outliers sufficiently far away from the median to skew our analysis?
* Are these outliers a result of input error or are they genuine values?

For our purposes, an outlier was:

**Less than (25th percentile – 1.5 IQR) or Greater than (75th percentile + 1.5 IQR)**

We will describe our post-hoc course of action in Section 3.4.

## Feature Generation and Transformation

To make the analysis useful, we decided to transform some numerical features into proportions.

|  |  |
| --- | --- |
| Columns | Columns description |
| FRP.Active/FPP | This ratio would tell us the risk appetite of the group, i.e., what proportion of the travelling group have taken the trip-cancellation insurance |
| FRP.Cancelled/FRP.Active | This ratio would tell us the proportion of the travellers who are insured and who ended up cancelling the trip. High value of this variable would be a red flag for the company |
| CancelledPAX/FPP | This ratio will tell us the measure of indecisiveness amongst the travelling group. Higher value of this variable would be a red flag for the company |
| FPP\_bin1 | 0 to 10 |
| FPP\_bin2 | 10 to 20 |
| FPP\_bin3 | 20 to 35 |
| NPS\_binary | binary score of 1/0 if avg 4 year NPS score > avg latest 3 year NPS score |

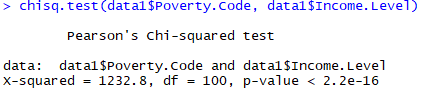
## Variable Selection

After the above steps, we went through all the variables, talking about their significance in the model and the economic theory that could give us an intuitive idea of how important we expected the variables to be, and why.

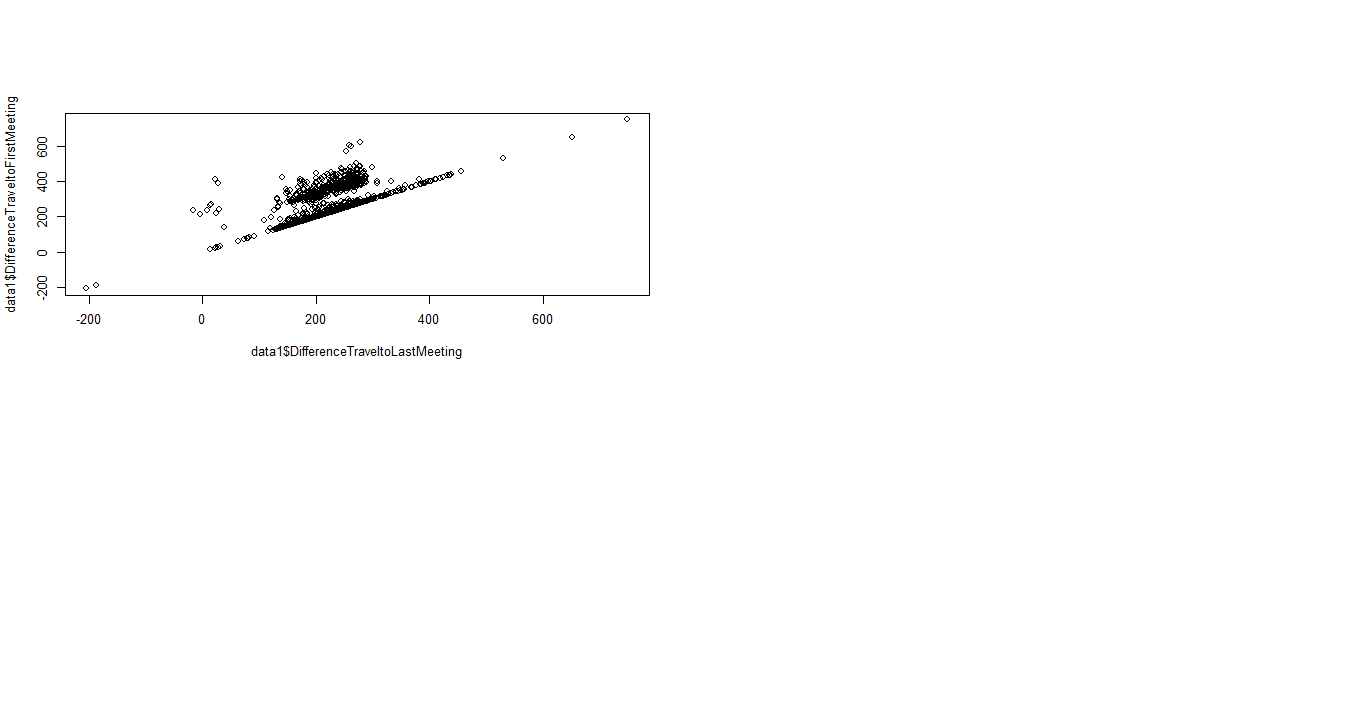
The objective was to come up with a baseline model using a subjective economic approach and then add or remove variables from it to improve it’s predictive power.

During this process, we faced the following challenges:

* **Related categorical variables:** When our understanding of the variables told us that the two variables were related, we verified our assumptions with a Chi-sq test before dropping one of the variables. In all the cases where we suspected two variables were related, the Chi-sq test agreed with us.



* **Related numerical variables:** Like the above, we decided to select one of many variables that we felt would be correlated. We verified our assumptions with either a scatter plot or a correlation test. This helped us mitigate multicollinearity even before we started the modelling process.



Eventually, we decided to go ahead with the following variables:

|  |  |
| --- | --- |
| **Columns** | **Columns description** |
| Is.Non.Annual. | 1/0 indicating if the group from this school typically skips a year in between programs. These will rarely repeat the very next year. |
| Days | The number of days the group was on the program and with one of the instructors. |
| Tuition | This is the price it costs each full-paying participant (FPP) to go on the program. West-coast air trips are more expensive per person than midwestern bus groups. |
| FRP.Active/FPP | This ratio would tell us the risk appetite of the group, i.e., what proportion of the travelling group have taken the trip-cancellation insurance |
| FRP.Cancelled/FRP.Active | This ratio would tell us the proportion of the travellers who are insured and who ended up cancelling the trip. High value of this variable would be a red flag for the company |
| Latest.RPL | This is the date that the last communication inviting people to join the group went out. Often this can be 6 to 9 months before the trip actually departed. |
| CancelledPAX/FPP | This ratio will tell us the measure of indecisiveness amongst the travelling group. Higher value of this variable would be a red flag for the company |
| School.Type | Public or private. |
| Parent.Meeting.Flag | 1/0 indicating whether a parent meeting was held. These are typically strong indicators of parent engagement and of a teacher who understands that these can be important to successfully organizing one of these out-of-school programs. |
| MDR.High.Grade | This is the highest grade in the originating school. |
| Income.Level | Like poverty code, an indication of ability of parents to pay for these programs. A is lowest, Q is highest, Z is unclassified. |
| School.Sponsor | This is an indication (1/0) of whether or not the school is officially sponsoring the trip. Mostly, though these programs draw from the same school, they are typically run independently. |
| SPR.New.Existing | EXISTING means that the group has traveled with STC before—most often the year before. NEW, with few exceptions, means that the school has never traveled before with STC. |
| FPP\_bin1 | 0 to 10 |
| FPP\_bin2 | 10 to 20 |
| FPP\_bin3 | 20 to 35 |
| NumberOfMeetingswithParents | Number of meetings with parents prior to the trip. |
| LastMeeting | The date of the last meeting with parents (NA if none held, may be same as the first meeting if only one meeting was held). |
| DifferenceTraveltoLastMeeting | The number of days from the last parent meeting to travel date. |
| SchoolGradeTypeHigh | The highest grade type in the school. |
| FPP.to.School.enrollment | The ratio of FPP to school enrollment. |
| FPP.to.PAX | The ratio of FPP to total PAX on the trip. |
| SchoolSizeIndicator | A label for the size of the school (S, M, L, S-M, M-L), by quintiles of sizes. |
| NPS\_binary | binary score of 1/0 if avg 4 year NPS score > avg latest 3 year NPS score |

## Imputation

At this point, we were able to impute the missing variables. Depending on the variables, we adopted one of the following strategies:

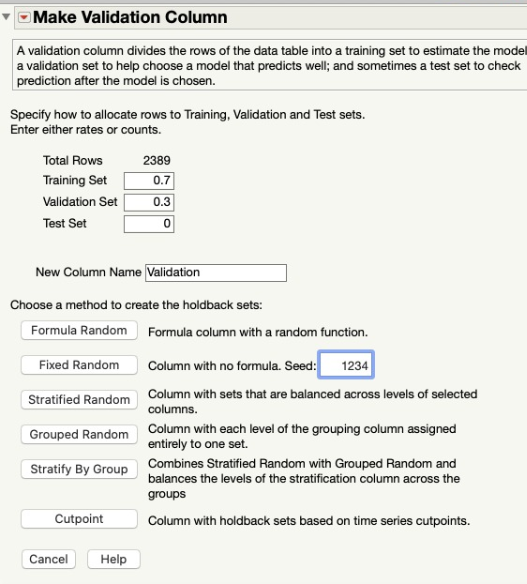
1. Dumb Imputation:
   1. Assumption that data is missing at random.
   2. Replace numerical missing value with mean of non-missing values.
   3. Replace categorical missing values with mode of non-missing values.
2. Smart Imputation:
   1. Does not assume all missing values are random and the same.
   2. Uses non-missing data to estimate missing values.

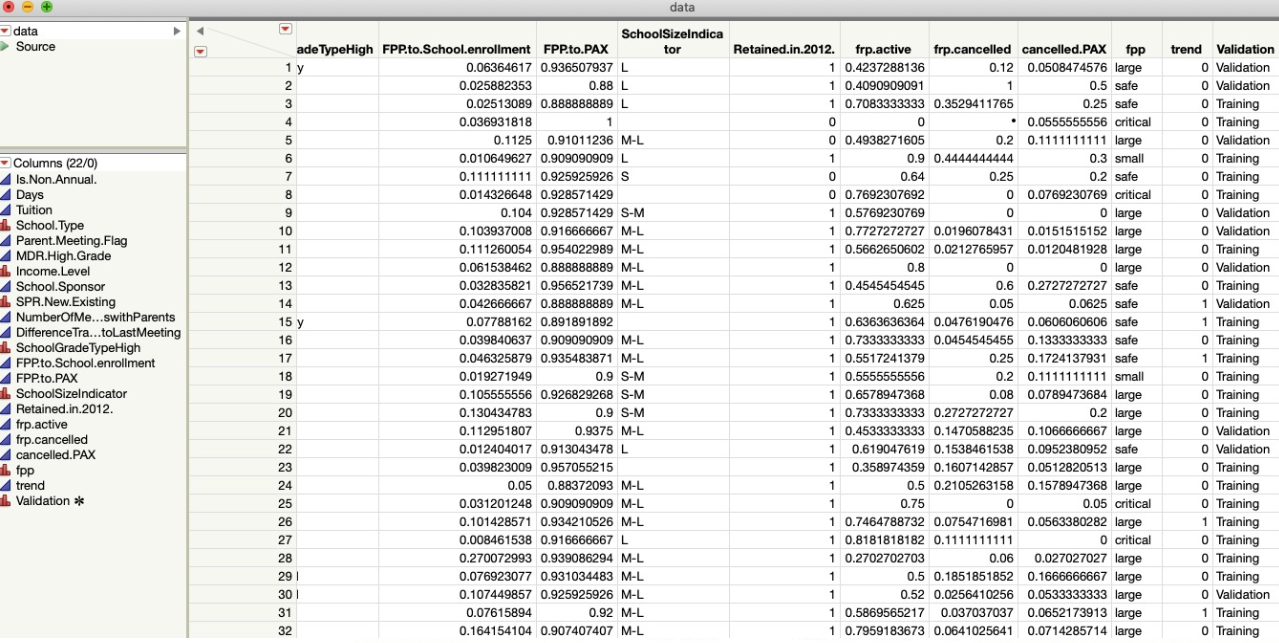
# predictive analytics

## Preparing the data for predictive analytics

In total, we have 2389 observations and 21 variables, including the response variable.

Before we begin applying machine learning models, we have divided the dataset into training and validation sets.



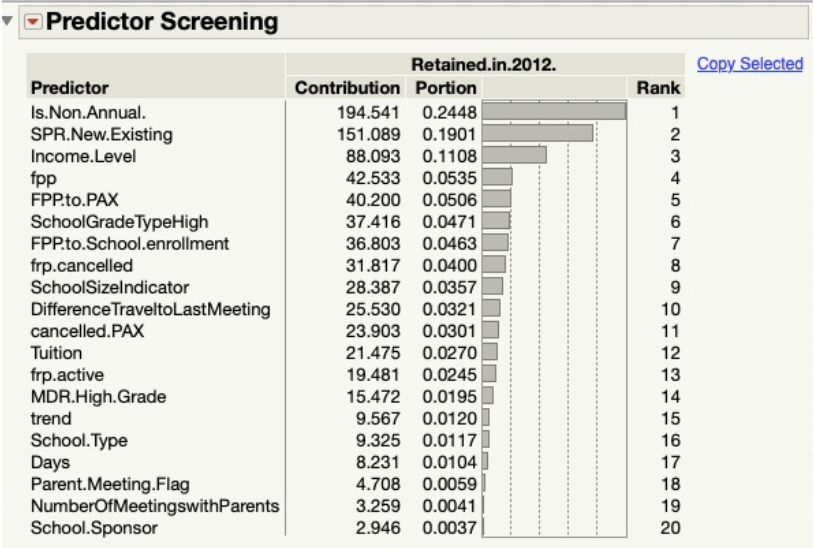


The training set containing 70% of the data and the validation containing 30% of the data, i.e., 1672 observations in the training set and the remining 717 datapoints in the validation set.

We have chosen fixed random division method, for splitting the data.

## Variable reduction

Since working with 20 odd predictor variables might give misleading results, we have decided to screen the most significant variables, to make the model simple and efficient.



We will be including the top 11 most contributing variables in building a logistic regression model, to predict the retention of a customer.

## model comparison

We have built multiple models, logistic regression model, neural network and classification tree amongst others. These three models have similar model accuracy parameters, which are listed below for comparison.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| S.No | Model | Accuracy | Misclassification Rate | AUC | Sensitivity | (1-Specificity) |
| 1 | Logistic Regression | 0.8 | 0.2 | 85.72 | 0.8589 | 0.2808 |
| 2 | Classification Tree | 0.79 | 0.21 | 84.7 | 0.8732 | 0.3427 |
| 3 | Neural Network | 0.79 | 0.21 | 84.86 | 0.8636 | 0.3218 |

We have selected logistic regression model out of the above three, as this model gives the best Sensitivity, with minimal (1-Specificity) value. This model also has lesser misclassification rate.

## Logistic Regression analysis

We have run the logistic regression on the categorical response variable “Retained in 2012”, including all the chosen predictor variables. We have selected 0.5 as the default threshold for classification model.

The significant variables effecting the response variable are,

Is.Non.Annual – The odds of a non-annual school getting retained with respect to that of an annual school is 0.04

SPR.New.Existing – If the customer is an existing customer, i.e., had already gone on trips organized by the company, their retention odds are 4.4816 when compared to a new customer

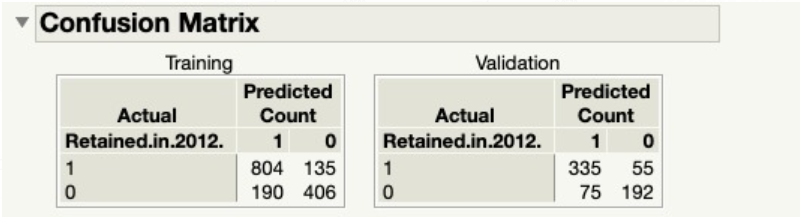
FPP – If the customer had sent a large group of students, more than 35, their retention odds for the next year are 3.77 compared to that of a group of less than 10 people. Else if the group of travelers is between 20 and 35, their retention odds for the next year are 2.33, compared to the same base group of less than 10 people.

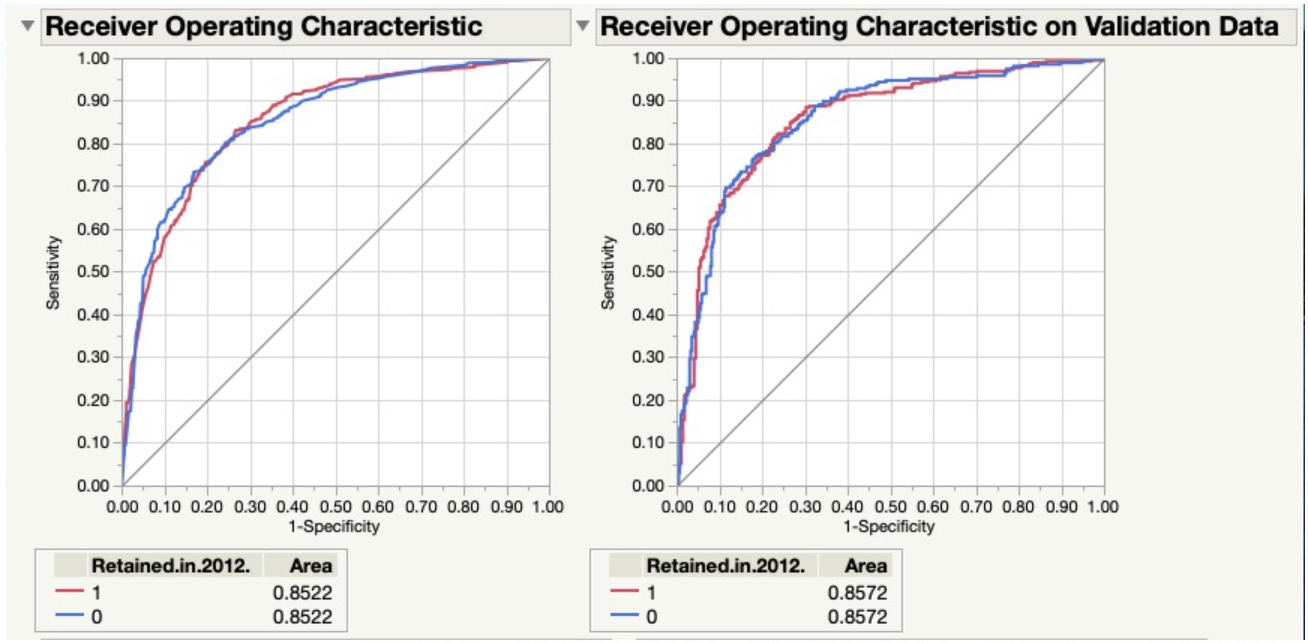
SchoolGradeTypeHigh – The schools which have grades till high school, are less likely to get retained for the upcoming year, with retention odds of 0.49 over the base group of schools, which are labelled as “undefined” in the company database.

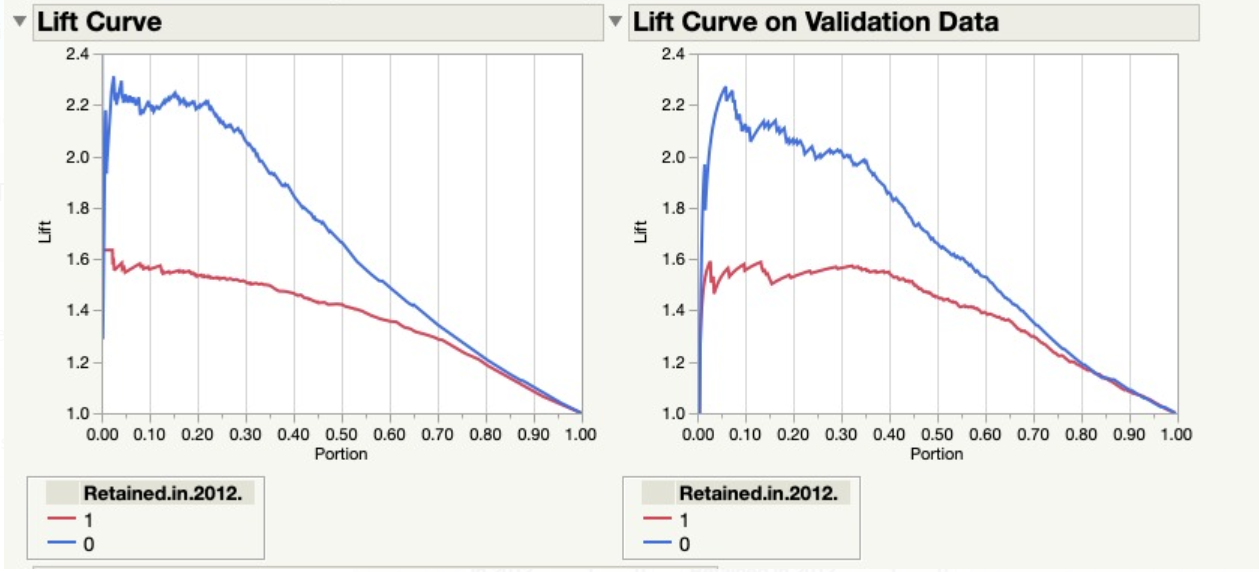
## Model Evaluation

To evaluate the performance of the model, we apply the model to the validation data set and calculate the following parameters.

* Sensitivity
* 1- Specificity
* Area under the curve
* Receiver Operating Characteristic
* Lift of the curve







The evaluation parameters are mentioned as follows.

|  |  |  |
| --- | --- | --- |
| Parameter | Training | Validation |
| Sensitivity | 0.85623 | 0.858974 |
| Specificity | 0.681208 | 0.719101 |
| (1-specificity) | 0.318792 | 0.280899 |
| Misclassification Rate | 21% | 20% |
| Accuracy | 79% | 80% |
| AUC | 0.8522 | 0.8572 |

## Assumptions behind building predictive model

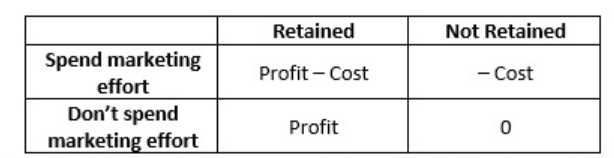
The following assumptions were made by us, before building the predictive model.

1. There are no other factors, apart from the ones listed in the master datasets, which influence the customers’ retention
2. The economic and social dynamics of travel and education industries remain mostly unchanged
3. The pricing policies of the company, like discounts or price hike, which may play a role in customer retention, are not taken into consideration.

# prescriptions

Look at the build model, we would like to suggest few prescriptive measures for the company.

1. Taking care of customers, who have a travel plan in their annual calendar is very important. These are the customers, if well taken care of, can be retained and the company can generate annual revenue.
2. The longer the customer has been retained by the company, the more likely it is that they will continue to be retained. These are the long-standing standing customers who are prospects for up-selling and cross-selling. The company can also rely on them for positive network effect.
3. The schools which have travel group size of more than 20 students are more likely to be retained. Therefore, the company needs to nurture these customers and try to work with the smaller size travel group schools, to bring them into the desired group size range.
4. Using an EV framework by incorporating the following cost matrix:



Use the above cost matrix with the probabilities of retention and non-retention as follows,



Then, sort the EV values in descending order, then take their cumulative sum and set a cutoff at the point which equals your marketing budget.

# Our Learnings

* Through this project, we had the hands-on learning experience of preparing the data for analysis. This stage included tasks like, converting the variables to required datatypes suitable for analysis, identifying outliers and skewness, imputing missing and NA values etc.
* We were exposed to critical model building stages like variable reduction, model selection and model evaluation.
* We have also learnt the aspects of machine learning, which would help in building models, to identify the underlying patterns in the historic data to predict the customer retention for the upcoming year.