Target Marketing And Cross Selling

- HARSH SAHAY

Problem Statement

Plumbing & Drain Service company offers a broad array of plumbing repair, sewer and drain cleaning services using its patented, proprietary machine. Demand Forecasting and customer churn is a challenging problem and for that it has collected records of historical data of their customers along with the kind of services they offered to them. The goal is to study the customer's data and to predict the kind of service to be offered for a customer based on his/her historical transactional data and churn prediction.

Data Insights

Performed basic data insights to get information about :-

The plumbing company’s customer base across the world.

Total number of historical transactions provided.

Total number of transactions pertaining to each customer type.

Total revenue made from each type of customer.

The trend in transactions made over the years 2011 to 2013 to get an initial understanding about the overall churn behavior.

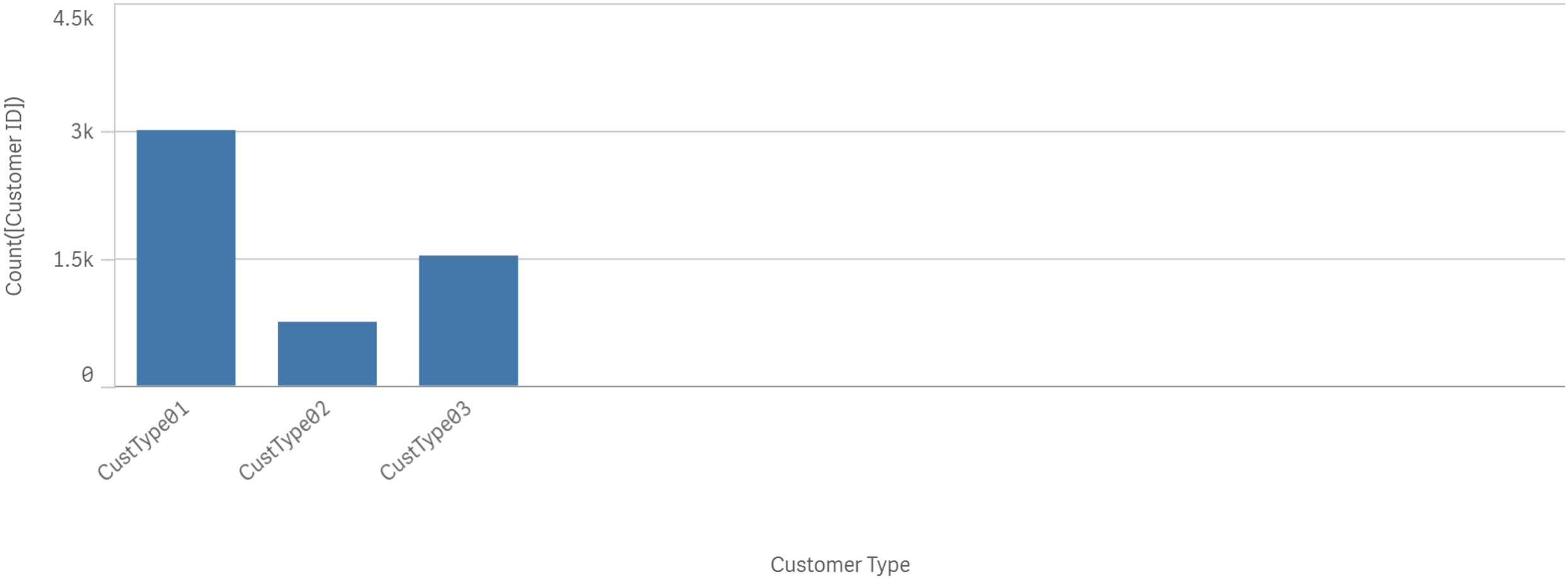
Tool Used :- Qlik Sense

Customer base of the plumbing company in various cities across the world

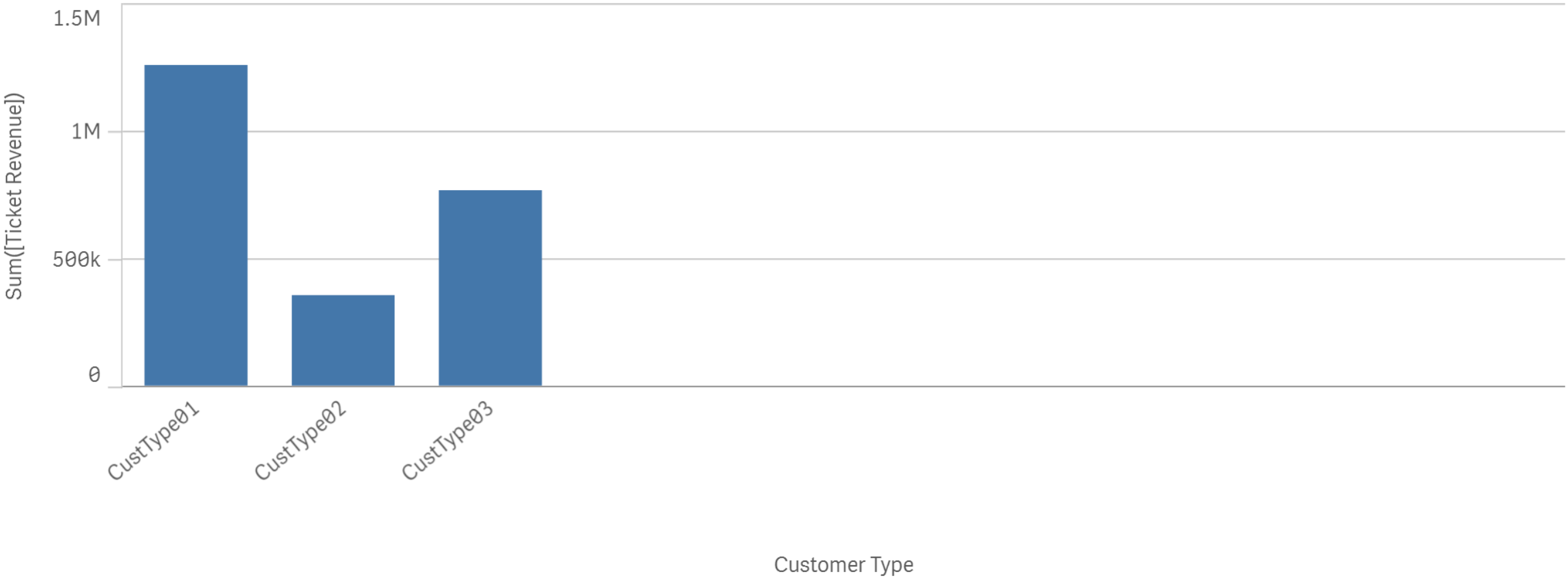
Total Number Of Transactions

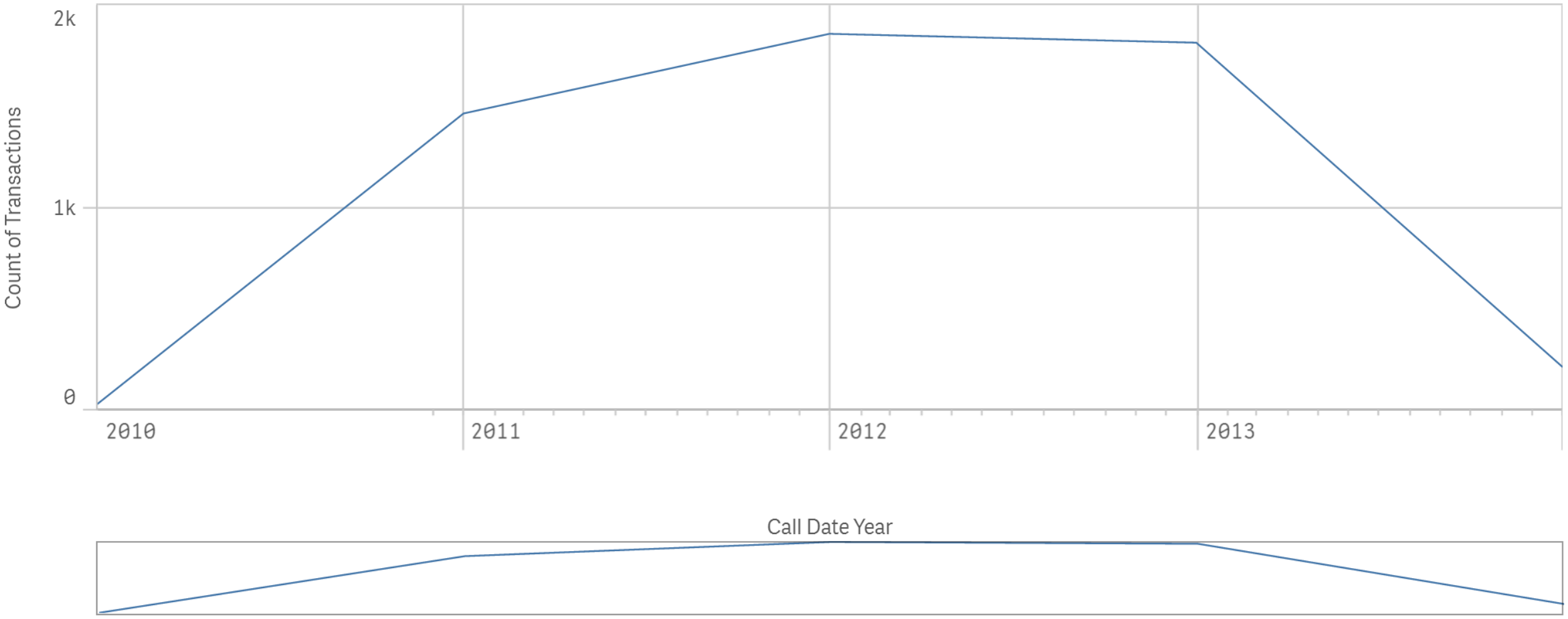
5.28k

Number of transactions pertaining to various customers belonging to each customer type



Revenue distribution for each customer type




2k
1k
0
Count of Transactions
2010
2011
2012
2013
Call Date Year

Initial Preprocessing

Required Attributes - "Customer.Type", "Customer.ID", "Job.Code", "Rev.Code","Call.Date","Setup.Date","Last.Service.Date","Complete.Date","Schedule.Date","Dispatch.Date“.

No missing values present in the data.

Factor Attributes - Rev.Code, Customer.Type, Job.Code.

Ordered the data w.r.t Customer.ID and Call.Date in ascending order.

Check the number of times each customer gets back. If a customer doesn't has enough historical data we drop that customer. **Assumption** -> Number of transactions for a customer must not be less than 3. Customers with ids C000197, C000436 have only two transaction records.

Feature Engineering

Number.Of.Days.Until.Next.Call -> Created an attribute to capture number of days after which a client calls again for next service.

Mean.Days -> For each customer-id add a new column to indicate the average number of days after which he/she comes back.

Dispatched.OnOrBefore.Scheduled.Date -> Added a new feature to calculate number of days between Scheduled and Dispatch dates. **Assumption** being that if Dispatch date is after Scheduled date the customer will churn.

Completed.OnOrBefore.Scheduled.Date -> Add a new feature to check if the Completion Date is on or before the Scheduled Date. Assumption being that if Completion date is after Scheduled date the customer is likely to churn.

Feature Engineering

Number.Of.Days.TO.Complete.Since.Call.Date -> Number of days taken by the plumbing company to complete the task since customer had called.

Completed.Within.2Days.Of.Call.Date -> If Complete.Date is more than 2 days of Call.Date. The customer is likely to churn. **Assumption** -> In majority of the transactions the Complete.Date is 2 days after the Call.Date.

Churn -> If(Number.Of.Days.Until.Next.Call > Mean.Days) OR
(Dispatched.OnOrBefore.Scheduled.Date == FALSE) OR

(Completed.OnOrBefore.Scheduled.Date == FALSE) OR

(Completed.Within.2Days.Of.Call.Date == FALSE)

Feature Engineering

Dropped the first transaction records for each customer-id for which Number.Of.Days.Until.Next.Call is 0.

Rev.Code.For.Non.Churn -> Added a new column which will indicate 0 if the customer has churned or indicate the Rev.Code if the customer has not churned.

Model Building for Churn Prediction

Used **Logistic Regression** and **CART**.

Logistic Regression :-

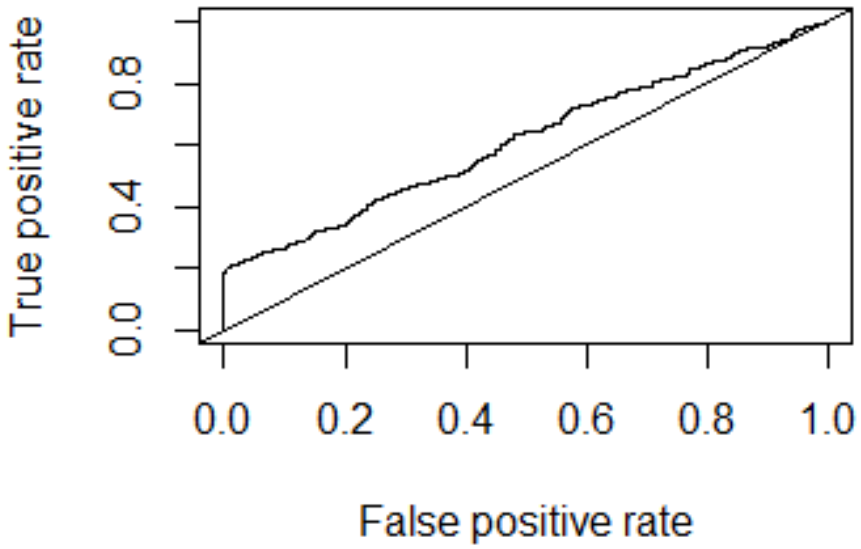
- **Challenges faced** -> The model did not converge initially and returned error. Used maxit = 100 as a parameter to glm. Maxit is the maximum number of iterations that the algorithm executes to get the best model.
- **Alias Co-efficients Problem** –> Two or more attributes infinitely correlated to each other. VIF test fails if any alias co-efficients are present. First removed the alias co-efficients and then perform VIF to check for multi-collinearity. Alias co-efficients were “Customer.ID”, “Setup.Date”, “Last.Service.Date”, “Mean.Days”, “Number.Of.Days.TO.Complete.Since.Call.Date”.

Model Building for Churn Prediction - GLM

Performance Metrices :-

- recall_Train = 67.64%
- recall_Test = 86.35%
- accuracy_Train = 65.56%
- accuracy_Test = 58.80%
- AUC = 61.37%

Model Building for Churn Prediction - GLM

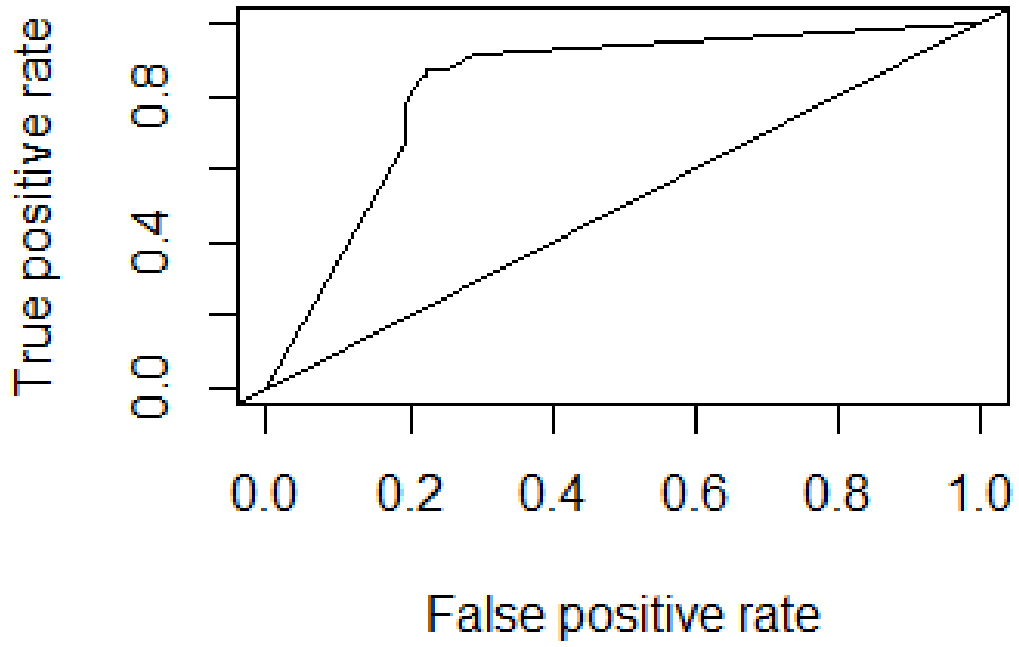


Model Building for Churn Prediction - CART

CART :-

- No challenges faced.
- No additional model specific pre-processing required on the initial pre-processed data.
- Performance Metrices :-
  - recall_Train = 96.24%
  - recall_Test = 87.47%
  - accuracy_Train = 96.81%
  - accuracy_Test = 83.52%
  - AUC = 82.78%

Model Building for Churn Prediction - CART



Model Building for Churn Prediction

With AUC = 82.78% CART is a better model for churn prediction as compared to GLM.

Model Building for Service Type Prediction

Used **C5.0** and **SVM**

- **C5.0 :-**
  - **Challenges Faced** :-
    - Execution in CART is very slow due to multi-level classification. Hence chose C5.0.
    - C5.0 doesn’t operate on data types 'POSIXct‘ and logical attributes(logi). Hence type-casted logi to factors and 'POSIXct’ to char
  - **Performance Metrices** :-
    - recall_Train -> 97.25%
    - recall_Test -> 97.25%
    - accuracy_Train -> 86.71%
    - accuracy_Test -> 86.71%

Model Building for Service Type Prediction

SVM :-

- **Challenges Faced :-**
  - Dropped POSIXct date attributes as SVM requires numeric attributes. The POSIXct attributes when type casted to numeric converts to a long literal which can be treated as id and not useful in prediction.

- **Kernel Used ->** Linear
- **Cost Value** -> 0.01 [Iteratively derived]
- **Gamma** -> 0.02 [1/Number of features]

Model Building for Service Type Prediction

Performance Metrices :-

- recall_Train = 98.70%
- recall_Test = 97.72%
- accu_Train = 91.40%
- accu_Test = 89.20%

SVM gives a higher accuracy on test-data as compared to C5.0.

Libraries Used

Package	Method Used
XLConnect	readWorksheetFromFile
plyr	arrange
car	vif
ROCR	prediction, performance
rpart	rpart
C50	C5.0
vegan	decostand
dummies	dummy
e1071	svm

References

http://stackoverflow.com/

Thank You