Uplifting revenue with controllable marketing variables in Telecommunication sector

1st Perera G.Y.V
209363M
University of Moratuwa
Colombo, Sri Lanka
pereragyv.20@uom.lk

2nd Weerasooriya P. A. A.
209392B
University of Moratuwa
Colombo, Sri Lanka
weerasooriyapaa.20@uom.lk

3rd Maldini P.A.N.
209355P
University of Moratuwa
Colombo, Sri Lanka
maldinipan.20@uom.lk

***Abstract*—This analysis is done in order to increase the revenue of a telecommunication company 'ABC' by offering promotions to its customers in a way such that the offered promotions maximize the net revenue uplift. The objective is to find out what customer attributes should be focused on for promotions, in order increase the mean revenue for a given month. The data set is explored using descriptive and diagnostics techniques. Predictive analysis is used to identify the target group of customers. Further analysis is done to distribute the given promotions among the selected target group within the budget.**

***Index Terms*—revenue, uplift, promotion, budget**

I. Introduction

To identify the most important attributes contributing to the revenue, descriptive and diagnostic analysis were done. Descriptive analysis revealed that revenue is highly correlated with data usage, data balance and recharge count, whilst moderately negatively correlated with the time since last recharge.

A linear regression model and a random forest model was built in order to predict the monthly revenue of the customers. Based on the models accuracy, the linear regression model was selected. Predictive analysis was done to estimate the expected monthly income for different scenarios consisting of selected predictors. A prescriptive analysis was conducted to find out how a promotional campaigns should allocate resources optimally in order to yield highest revenue over allocated resources.

II. Methodolody and results

A. Data collection

The data set contains the details of 400 prepaid users of a leading global telecommunication company. It has 38 attributes such as; data usage, talk time, recharge amount for a month. The data were collected in 2018

B. Data cleaning up

The data set was checked for issues by checking frequencies to identify unreasonable distributions, unreasonable values, and misinformative values. For categorical attributes, possible values were identified and evaluated by checking the unique values in each. Also, types of numeric values were checked.

One missing data point in 'district' was replaced by the mode 'COL'.

There were no other issues with the data set. However, some dichotomous attributes were mapped to 0 and 1 for the analysis.

C. Descriptive and Diagnostic analysis

1) Descriptive analysis: Central tendency and dispersion measurements were calculated for each attribute of the data set. Further, following were analyzed:

- Revenue by district and language: Sinhala was observed to be the prominent language except for Ampara, Nuwara-Eliya, Mathale, Batticaloa, Kalutara, Jaffna, Puttalam. Also, the majority of the customers were from Colombo district. (Fig. 1)
- Customer counts by Dual Sim flag: Majority of the customers (95.25%) use dual sim facilities. (Fig. 2)
- Customer counts by Smart Phone flag: Majority of the customers (57.5%) use smartphones. (Fig. 3)
- Revenue distribution:
  - mean: 296.995975
  - std: 386.290629
  - min: -0.350000
  - 25%: 18.070000
  - 50%: 148.590000
  - 75%: 395.692500
  - max: 1999.830000

  (Fig. 4)
- Revenue by district: Highest revenue is generated from Colombo, followed by Gampaha and Kandy. (Fig. 5)
- Data Usage vs. Revenue: Data usage and revenue is positively correlated. However, there is a set of customers who use almost no data, but still contributing to revenue. This implies that they are using other services such as calls.(Fig. 6)
- Call Duration vs. Revenue: Data usage and revenue is positively correlated. However, there is a set of customers who use almost no call time, but still contributing to revenue. This implies that they are using other services such as data.(Fig. 7)

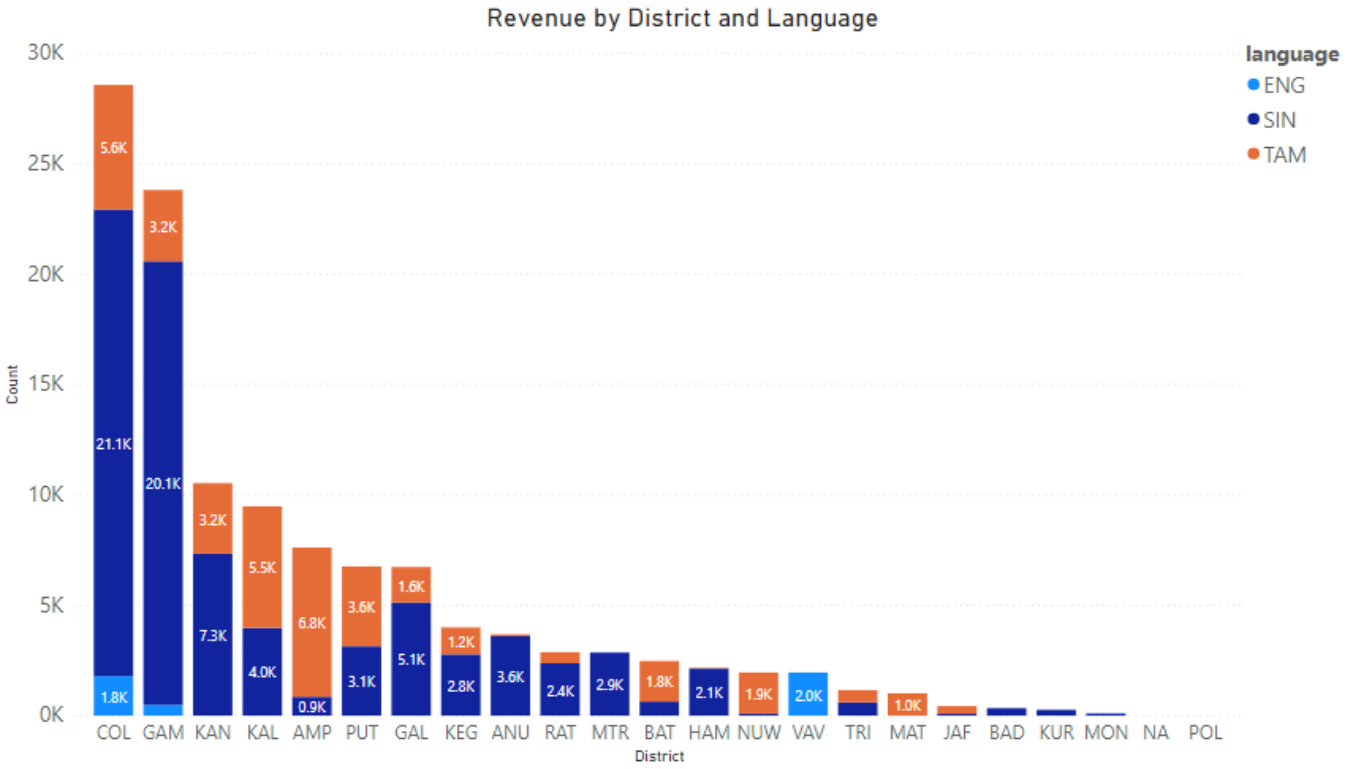

Fig. 1. Revenue by district and language.

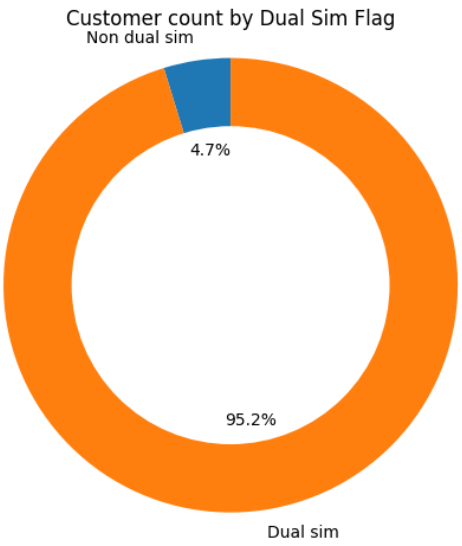

Fig. 2. Customer counts by Dual Sim flag.

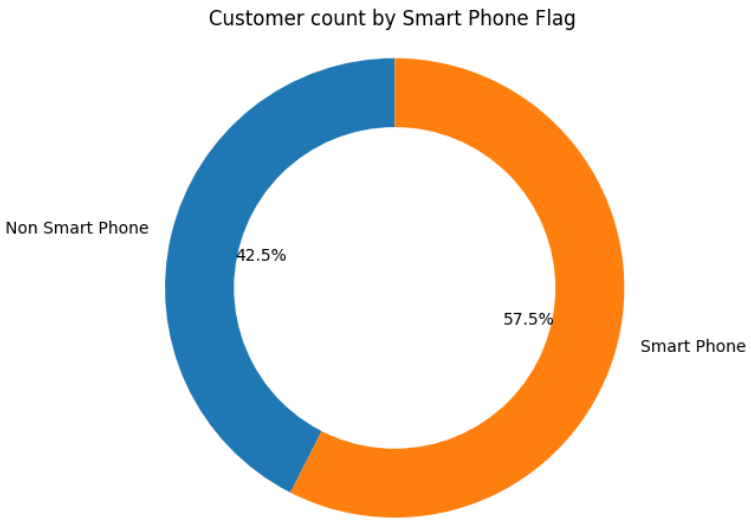

Fig. 3. Customer counts by Smart Phone flag.

- Call Duration vs. Data Usage: Its observed that there are two sets of customers, one group uses data but almost no call time, the other uses call time but almost no data. This may be due to two products, such as data package and a call package (Fig. 8)

2) : Diagnostic analysis

- Highly correlated variables for revenue-rs : data-mb, data-balance, recharge-count
- time-since-last-recharge and revenue-rs variables show moderate negative correlation.

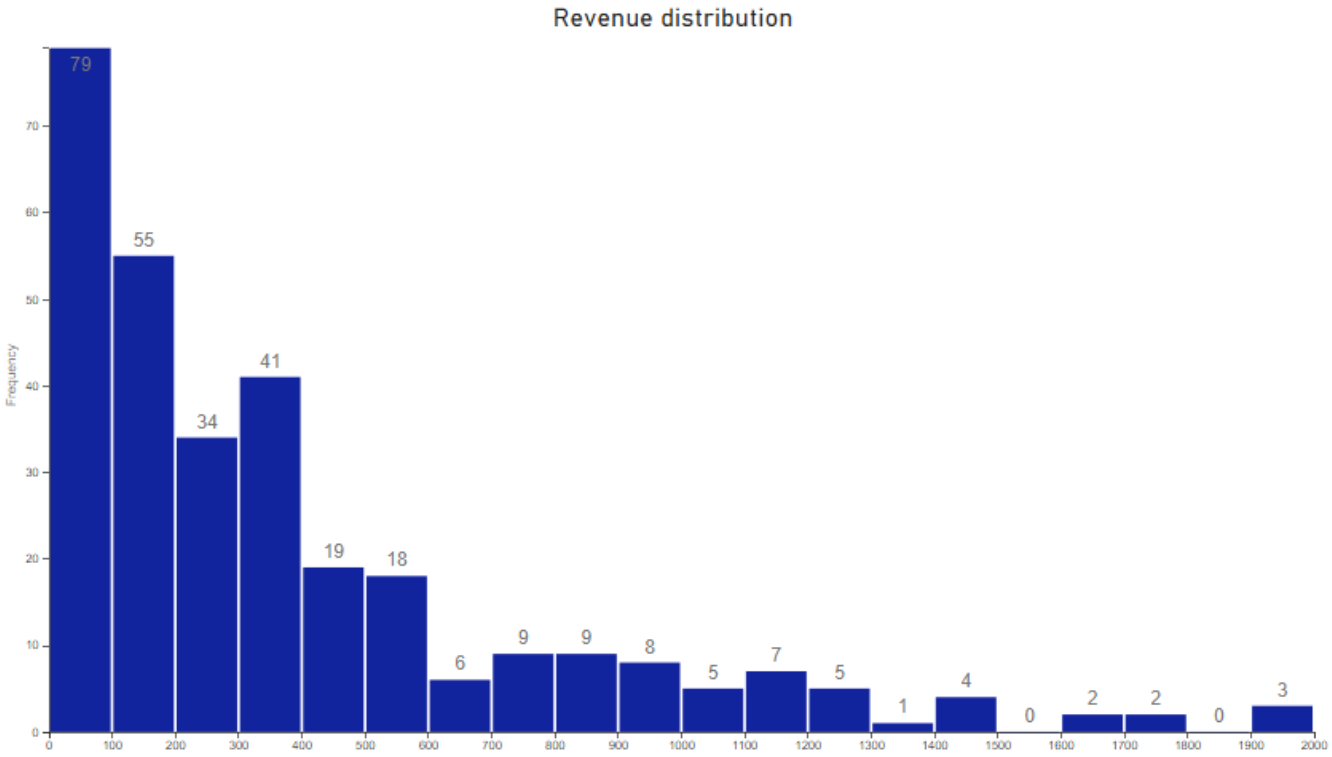

Fig. 4. Revenue distribution.

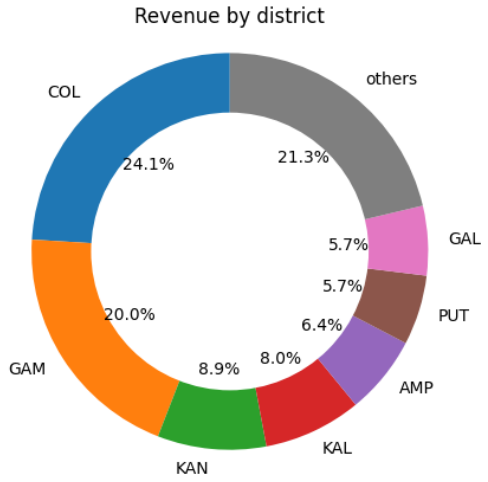

Fig. 5. Revenue by district.

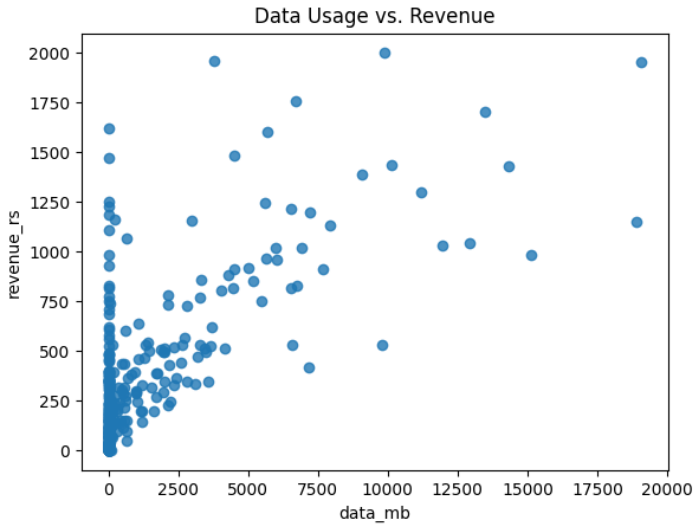

Fig. 6. Data Usage vs. Revenue.

- mean 4.630000
- std 13.722933
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 1.000000
- max 61.000000

- An average customer stays 5 days without any activity in the network.
- 75% of the customers have a very less number of days since their last activity in the network. (Fig. 10)

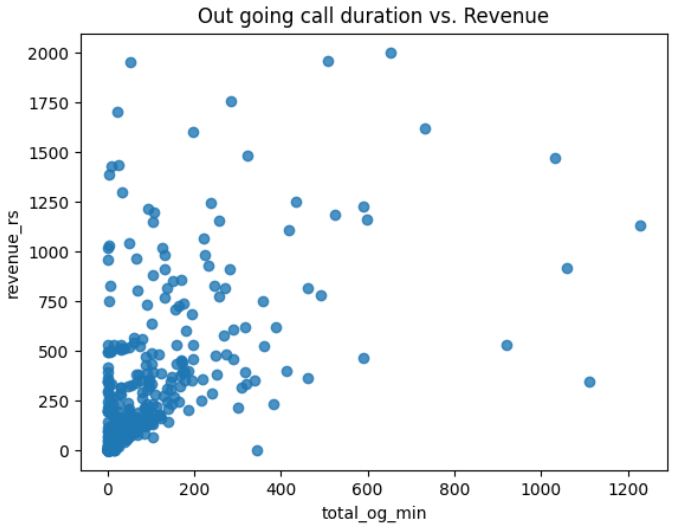

Fig. 7. Call duration vs. Revenue.

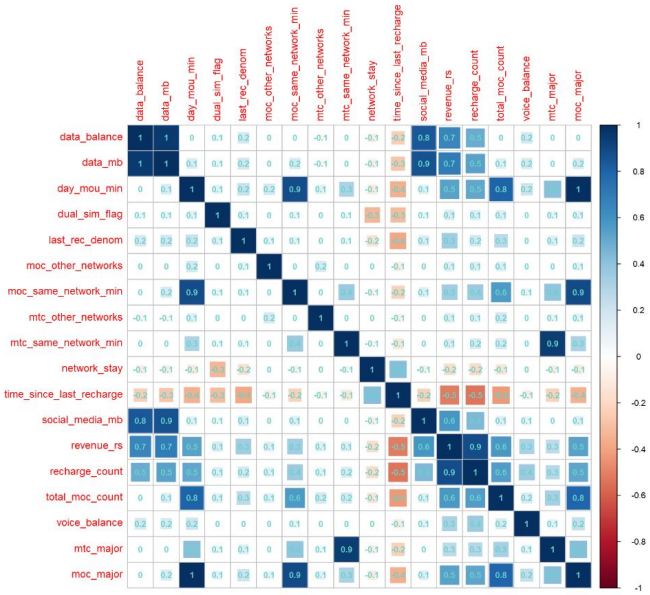

Fig. 9. Correlation Plot

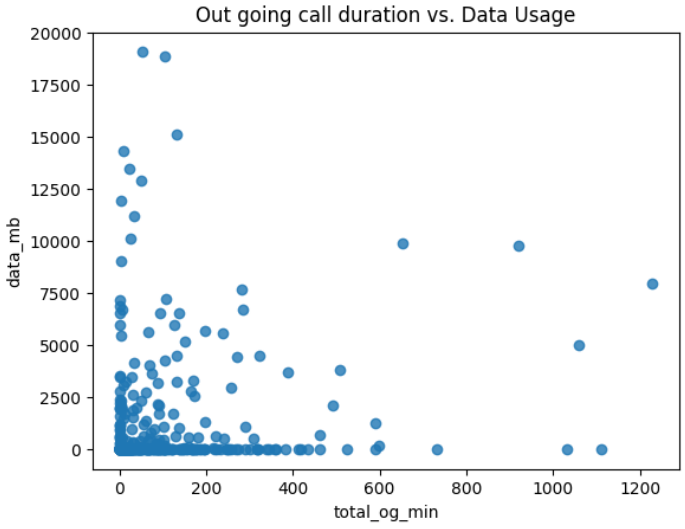

Fig. 8. Call Duration vs. Data Usage.

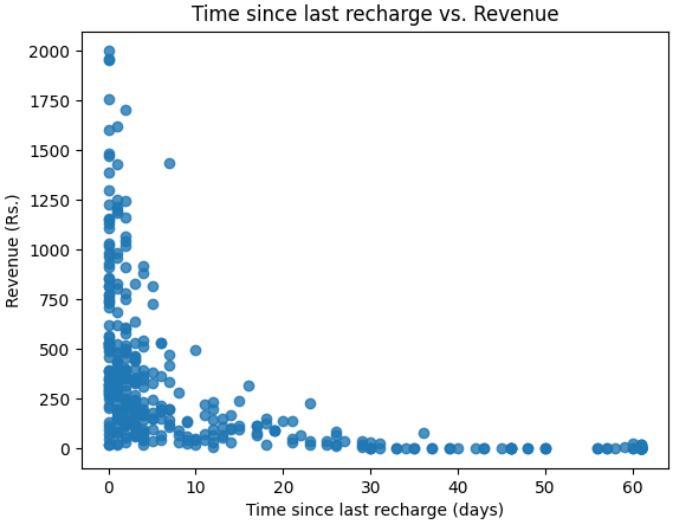

Fig. 10. Time since last recharge vs. Revenue.

3) : Predictive analysis

A regression analysis was done in order to predict the monthly revenue of each customer. Both random forest model and linear regression model were built and the linear regression model was selected based on the accuracy measures.

Multicollinearity is a problem when building a regression model because it can increase the variance of the coefficient estimates and make the coefficient estimates unstable and difficult to interpret [2]. Hence, the feature selection was done in two parts. First, the highly correlated variables were identified and then feature importance was calculated using a base random forest model. Top 15 features were selected for model building by considering the correlation and importance.

- Highly correlated variables :
  The highly correlated variables were identified as follows. Pearson correlation was used to calculate the correlation coefficients and the variable sets with coefficient > 0.80 were considered as 'high'.

TABLE I
CORRELATED VARIABLES

Variable 1	Variable 2	Correlation
total og min	moc major	0.9970
total og min	day mou min	0.9957
day mou min	moc major	0.9922
revenue rs	recharge value	0.9886
data balance	data mb	0.9545
mtc same network min	mtc major	0.9460
time since last activity	network stay	0.9391
data revenue	data mb	0.9329
data revenue	data balance	0.9119
moc same network min	moc major	0.8967
recharge count	recharge value	0.8905
total og min	moc same network min	0.8884
revenue rs	recharge count	0.8876
day mou min	moc same network min	0.8829
social media mb	data mb	0.8669
social media mb	data balance	0.8428
day mou min	total moc count	0.8051

- Feature importance :
  A base random forest model was built with 10000 trees. Feature importance measures were obtained as follows.

TABLE II
FEATURE IMPORTANCE

Variable	Importance
recharge value	0.9758
data mb	0.0018
rc slab 30	0.0016
social media mb	0.0012
recharge count	0.0011
time since last recharge	0.0009
voice balance	0.0009
data balance	0.0008
network stay	0.0008
mtc idd min	0.0008

[a]Only the first 10 variables are shown here

- Selected variables :
  - recharge value
  - data mb
  - rc slab 30
  - time since last recharge
  - voice balance
  - data balance
  - network stay
  - mtc idd min moc same network min
  - mtc same network min
  - moc idd min
  - total moc count
  - mtc other networks
- Model building :
  A linear regression model was built to predict the monthly revenue of the customers. 80% of the data was used as the training data set. The mean squared error, which is the average squared difference between the estimated values and the actual value was used on testing data set (the 20% left) to evaluate the model results.
  The best model results:
  - Mean Absolute Error : 17.30
  - Mean Squared Error : 801.08
  - Root Mean Squared Error : 28.30

TABLE III
REGRESSION EQUATION

Variable	coefficient
recharge value	0.7166
data mb	0.0146
voice balance	- 0.1195
data balance	- 0.0013
moc same network min	0.0981
day mou min	- 0.1660
time since last recharge	- 0.1040
intercept	7.5143

The recharge value, data mb and moc same network min has positive effect on the response variable revenue. Voice balance, data balance, day mou min and time since last recharge has positive effect on the response variable revenue.

The monthly revenue was predicted for the group of customers in the testing data set.

Actual total revenue = Rs.19690.77

Predicted total revenue = Rs.19756.54

4) : Prescriptive analysis

Actionable insights were given by analysing the results of the descriptive, diagnostics and predictive analyses.

In tele-communication sector, the usual approach for up sell-cross sell is sending out the same promotion for the entire customer base. As the usage and network stay is different from customer to customer, this approach has limitations in maximizing the net revenue uplift.

The main focus in this analysis was to optimize the promotion assignment in order to maximize the net revenue uplift.

As the data set was only for a month, the testing set was used as the experimental data set for the next month by

assuming that the monthly revenue stays the same for the next month without any promotion campaign.

- Promotion related details
  - Voice promotion
    - Internal cost : Rs.5
    - Mapped controllable predictor : moc same network min
    - Details : Recharge Rs.50 and talk for Rs.100 in the same network ($\approx 65\ \ minutes$)
  - Data promotion
    - Internal cost : Rs.10
    - Mapped controllable predictor : data mb
    - Details : Recharge Rs.49 (800 mb) and use 1GB
- Revenue prediction with the controllable variables

  e.g: Offer the 'voice promotion' to customer 'A' and the 'data promotion' to customer 'B'

  The controllable predictor 'moc same network min' value of the customer 'A' is increased by the promotion amount (65 minutes) when the voice promotion is offered to the customer 'A'.

  The controllable predictor 'data mb' value of the customer 'B' is increased by the promotion amount (1000 mb) when the data promotion is offered to the customer 'B' (Assuming that the customer 'A' and 'B' accept the promotion if and only if he/she has non zero historical values for the corresponding variable)

  Before promotion assignment :

customer	data mb	moc same network min
A	950	60
B	400	90

  After promotion assignment :

customer	data mb	moc same network min
A	950	125
B	1400	90

  The revenue prediction after the promotion campaign is done by feeding the updated predictor values in to the linear regression model which was explained in the predictive analysis. The revenue uplift can be calculated by taking the difference of the total predicted revenue before and after the promotion campaign.
- Promotion assignment

  The experimental set-up was done by identifying the needy customers to accept the respective promotion. The variables 'moc same network min' and 'data mb' were scaled using MinMaxScaler. The respective promotion is assigned to the customers whose scaled value is greater than a threshold ($\alpha = 0.02$)

The cost is calculated as below :

No of customers with the data promotion = 18

No of customers with the voice promotion = 27

Cost for both promotions = Rs.10 x 18 + Rs.5 x 27

Total internal cost = Rs.315

The predictors 'moc same network min' and 'data mb' were increased and the monthly revenue prediction was done by feeding the updated predictor values in to the linear regression model.

In the experimental set-up :

- Predicted total revenue = Rs.23573.22
- The revenue uplift = Rs.21791.50 - Rs.19756.54
  = Rs.2034.97
- Net revenue uplift = The revenue uplift - Total internal cost
  = Rs.2034.97 - Rs.315.00
  = Rs.1719.97

III. Conclusion

This analysis suggests a method of assigning the promotions while tuning the predicted figure of the revenue uplift. The promotion assignment can be done using the same model by varying the threshold α.

References

[1] Kim, Y. S., et al. Churn management optimization with controllable marketing variables and associated management costs. Expert Systems with Applications (2012), http://dx.doi.org/10.1016/j.eswa.2012.10.043

[2] Editor, M., 2020. What Are The Effects Of Multicollinearity And When Can I Ignore Them?. [online] Blog.minitab.com. Available at: ¡https://blog.minitab.com/blog/adventures-in-statistics-2/what-are-the-effects-of-multicollinearity-and-when-can-i-ignore-them¿ [Accessed 10 April 2020].

[3] Medium. 2020. A Beginner'S Guide To Linear Regression In Python With Scikit-Learn. [online] Available at: ¡https://towardsdatascience.com/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-83a8f7ae2b4f¿ [Accessed 10 April 2020].

[4] Medium. 2020. Explaining Feature Importance By Example Of A Random Forest. [online] Available at: ¡https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e¿ [Accessed 10 April 2020].