

## Compete ▾

### Talaria Forecasting - Financial Forecast BB Sandesh Brand 4 - Refinement

Code **TC021** Machine Learning Python  
Data Science Linux

[Recommended Challenges](#)

[Recommended THRIVE Articles](#)

#### Key Information

1st	2nd
<b>\$750</b>	<b>\$350</b>

[Unregister](#)

[Submit](#)

Next Deadline: **Submission** | 23h 50mins until current deadline ends

[Show Deadlines](#) ▾

[DETAILS](#)

[REGISTRANTS \(45\)](#)

[SUBMISSIONS \(5\)](#)

[CHALLENGE FORUM](#)

## Challenge Overview

### Prize

1st place - \$750

2nd place - \$350



### Background

Over the last few months, a series of challenges have been run to generate a series of high quality financial forecasts for a consumer Broadband and Mobile provider, 'Sandesh'. These challenges have been known as 'CFO Forecasting' in various formats. As a result, a high quality, high accuracy series of algo

have been produced for a number of financial target variables.

[Support](#)

This challenge is being rerun to generate a similar high quality forecast for BB Sandesh Brand 4, a 'sister brand' owned by the same client, offering a similar suite of Broadband and Mobile products to the Small and Medium Enterprise market ('SME').

### Challenge Objective

The objective of this challenge is to generate the highest accuracy predictions possible for the four variables outlined below. The accuracy of the forecast must at least improve on the Threshold target quoted for each variable/product.

- The model should be tailored to a 12-month forecast horizon but must be extendable beyond this time period. The accuracy of a prediction will be evaluated using MAPE (Mean Average Percentage Error) on the privatized data set.
- **For robustness calculation** do the forecast in iteration as explained in section: *Quantitative Scoring point 2*.
- The Prediction window for **MAPE calculation is Oct19 – March20. Training should not be done on the prediction window data points.**

### Business Context

Following are the details about the product and target variables of Sandesh Brand 4 which need to be forecasted.

#### Broadband SME

- Product:
  - **Falcon** – Sandesh's main broadband product (reaching maturity, available in most of the country, faster download speeds than the legacy product)
- Target Variables:
  - **Average Revenue per New Customer** – the average monthly revenue paid by new Falcon customers in the first month of the customer's contract
  - **Average Revenue per Existing Customer** – the average monthly revenue paid by all subscribers in the Falcon customer base per month for the service



- **Gross Adds(Norm)** – *the number of new subscribers to each individual product joining the brand during a month*
- **Net Migrations (Norm)** - *The number of subscribers who remained with the brand but moved to another product. Usually and upgrade to faster broadband speed.*

Gross Adds (Norm) and Net Migrations (Norm) are seasonal and discontinuous variables and vary significantly from month to month depending on competitor pressure at that point in time. ARPU New & Existing are continuous variables.

#### Challenge Thresholds and Targets

Broadband Sandesh Brand 4			
		Threshold	Target
Average revenue per <b>existing</b> customer	Falcon	450%	300%
Average revenue per <b>new</b> customer	Falcon	10%	5%
Gross Adds(Norm)	Falcon	500%	330%
Net Migrations(Norm)	Falcon	25%	18%

#### Outlier Treatment

Within the data sets provided there are a number of outliers due to reasons outside of the market. These outliers should be treated as deemed appropriate to improve forecasting performance.

Any such treatment must be clearly documented and explained in the submission. Here are some examples that should be considered, though these are not exhaustive.

#### Average Revenue per New Customer





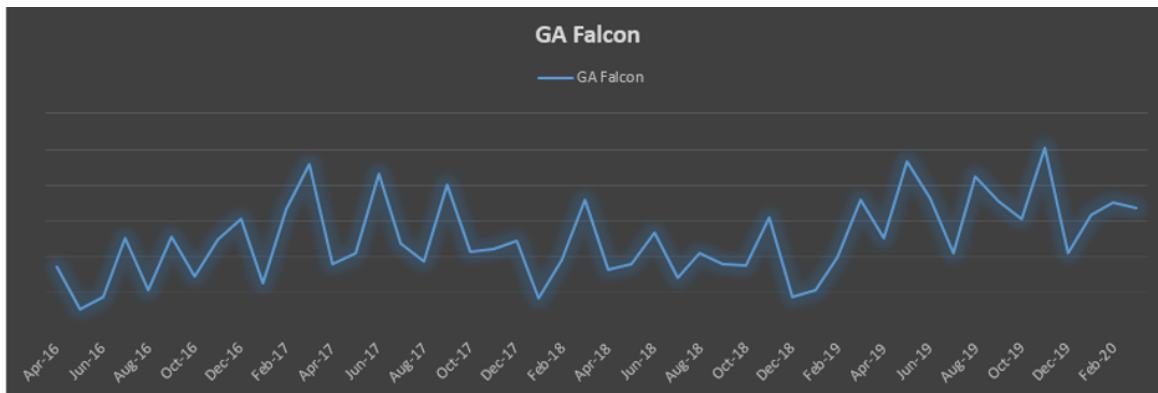
#### Outlier adjustment required:

There is a significant 'step change outlier' between Mar 18 - Apr 18 and Apr19 – June19. This is due to an artificial re-correction and should be adjusted. Please build into the submission; clearly document how the 'step change outlier' has been handled in your model. The treatment needs to be generic so that it can be applied to real data set.



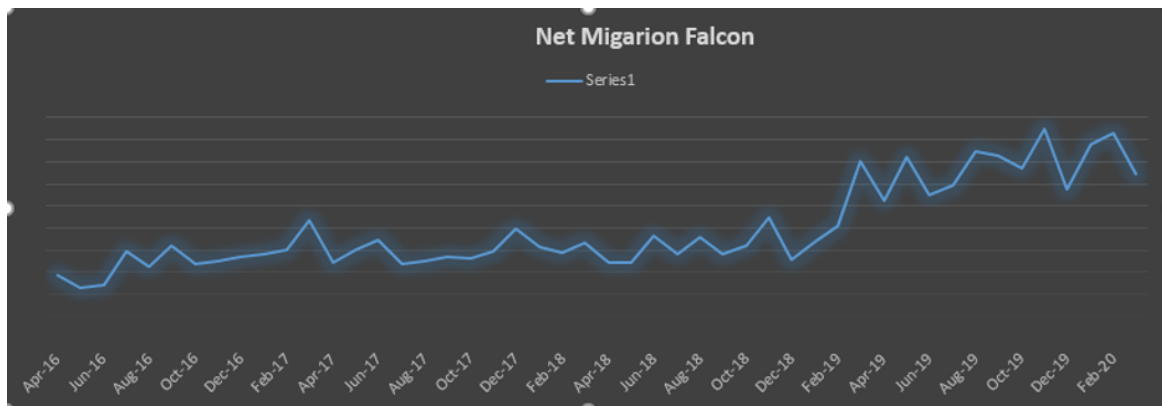
There are a couple of apparent outliers in the data sets. For the purposes of this challenge, outliers need to be identified and treated to improve the forecast of the resulting data.

### Gross Add Falcon



We have achieved MAPE of 12% on actual data. What we observed that starting from April-19 there is a mean shift of the values. Also for many months the surge and drops are not look alike the historical months.

### Net Migration Falcon



For this variable we have achieved MAPE of 14% on the real data. Starting from January-19 There is a uptrend of the data seen.

### Financial Year Modeling:

Sandesh reports its financial year from April - March. This may contribute to seasonality based on financial year, and quarters (Jun, Sep, Dec, and Mar), rather than calendar year.

### Anonymised and Privatised Dataset:

'Z-score' is used to privatise the real data.

For all the variables, following is the formula used to privatise the data:

$$z_i = (x_i - \mu) / \sigma$$

where  $z_i$  = z-score of the  $i$ th value for the given variable

$x_i$  = actual value

$\mu$  = mean of the given variable

$\sigma$  = standard deviation for the given variable



## Modeling Insight Derived from Previous Challenges

For Net Migration (Norm) and Gross Add (Norm) we have previously used linear regression, which has outperformed ARIMA, SARIMA and LSTM. But the MAPE was more than 10% in real data. We are trying to achieve nearly 7% MAPE on the Real data.

For ARPU New User we have tried with SARIMA model  $(0,1,0)(0,0,1,12)$  and could reach to 10.5% MAPE on real data. We have tried with log and Box-cox conversion on the raw data. Here also we are targeting nearly 7% of MAPE on the real data.

For ARPU existing user, please note that after Feb2019 the trend is purely downward. We have tried with ARIMA and it produced 5% MAPE. We are trying to minimize it.

## Final Submission Guidelines

### Submission Format

Your submission must include the following items

- One prediction file over the forecast time period (Oct19 – March20). We will evaluate the results quantitatively (See below)
  - Please use Time Period, Generic LookupKeys as the column names.
  - The values in Time Period column are something like 2019-08
  - The values in each Generic LookupKey column are the predicted values, i.e., floating numbers.
  - The final spreadsheet has a  $N \times (M+1)$  shape, where N is the number of time periods and M is the number of variables that we want to predict in this challenge. "+1" is for the Time Period column.
- Six prediction files for the robustness purpose following the same format correspond to 6 sliding windows (mentioned in the robustness definition).
- A report about your model, including data analysis, model details, local cross validation results, and variable importance.
- A deployment instruction about how to install required libs and how to run.



## Expected in Submission

- 1 **Working Python code** which works on the different sets of data in the same format
  - .
- 2 Report with clear explanation of all the steps taken to solve the challenge (refer section “Challenge Details”) and on how to run the code
  - .
- 3 **No hardcoding** (e.g., column names, possible values of each column, ...) in the code is allowed. We will
  - . run the code on some different datasets
- 4 All models in one code with clear inline comments
  - .
- 5 Flexibility to extend the code to forecast for additional months
  - .

## Quantitative Scoring

### First, MAPE on Prediction Window

Given two values, one ground truth value (**gt**) and one predicted value (**pred**), we define the relative error as:

$$\text{MAPE}(\text{gt}, \text{pred}) = |\text{gt} - \text{pred}| / \text{gt}$$

We then compute the **raw\_score(gt, pred)** as

$$\text{raw\_score}(\text{gt}, \text{pred}) = \max\{0, 1 - \text{MAPE}(\text{gt}, \text{pred})\}$$

That is, if the relative error exceeds 100%, you will receive a zero score in this case.

The final MAPE score for each variable is computed based on the average of **raw\_score**, and then multiplied by 100.





**Final score = 100 \* average( raw\_score(gt, pred) )**

MAPE scores will be 50% of the total scoring.

You will also receive a score between 0 and 1 for all the thresholds and targets that you achieve. Each threshold will be worth 0.03 points and each target will be worth 0.03 points. If you achieve the target for a particular variable you'll get the threshold points as well so you'll receive 0.07 points for that variable. Your points for all the variables will be added together.

## Second, Robustness

Once model building is done, robustness of the model is to be calculated. For this we need to do evaluation on a rolling forecasting from origin. See the below image for understanding. (This is just a sample image in actual, dates will be changed for the training and forecasting horizon)

Iteration Number	Training Window	Forecast Window	File name in robust dir
1	April14 <-----Training Window Iteration1-----> March19	April19<-----Forecast period---->March20	/robust/submission1.csv
2	Apr-14 Feb-19 Mar-19	Feb-20	/robust/submission2.csv
3	Apr-14 Jan-19 Feb-19	Jan-20	/robust/submission3.csv
4	Apr-14 Dec-18 Jan-19	Dec-19	/robust/submission4.csv
5	Apr-14 Nov-18 Dec-18	Nov-19	/robust/submission5.csv
6	Apr-14 Oct-18 Nov-18	Oct-19	/robust/submission6.csv

1 Every horizontal line represents one iteration.

.

2 Blue windows are the training period and Orange windows are Forecasting period.

.

3 One separate directory to be created with the name "robust" to store the forecast for all the iterations.

.

4 Forecast to be saved in a .csv file with the name submission affixed with the iteration number.

.

5 Separate python function/module to be created, which will call the model to generate the forecast for different periods as explained in the above image. This is required for code modularity.

6 The function/module for the robustness, should have below input parameters.

.

1 Start date of the training window of the iteration 1. (April2019)

.

2 End Date of the training window of the iteration 1. (March 2020)



- 2 End Date of the training window of the iteration 1. (*March-2020*)
- .
- 3 Forecast period i.e number of months to forecasts. (*12 months*)
- .
- 4 Number of iteration. (*12 iterations to be done*)
- .
- 5 For subsequent iteration Train/Forecast start and end month should be automatically calculated
  - . based on the input given in step A and B as shown in the above image.
- 7 While running this module, you should use the final model based on training data till September 2019.
  - . For an example if it's an ARIMA model then p,d,q values should be the same throughout all the iteration. If it's a LSTM then hyper param such as epochs, number of LSTM units, look back/look forward etc should be the same as your final model built for date range mentioned in point 8. All the iterations should run on the same parameter configuration.
- 8 Robustness will be calculated based on MAPE over 12 iterations.
- .

### Judging Criteria

Your solution will be evaluated in a hybrid of quantitative and qualitative way.

- Effectiveness (80%)
  - We will evaluate your forecasts by comparing it to the ground truth data. Please check the "Quantitative Scoring" section for details.
  - The smaller MAPE the better.
  - Please review the targets and thresholds above as these will be included in the scoring.
  - Along with MAPE, we will calculate the robustness (at our end) as explained in *Quantitative Scoring point 2* section.
- Clarity (10%)
  - The model is clearly described, with reasonable justifications about the choice.
- Reproducibility (10%)
  - The results must be reproducible. We understand that there might be some randomness for ML models, but please try your best to keep the results the same or at least similar across



different runs.

## Payments

Topcoder will compensate members in accordance with our standard payment policies, unless otherwise specified in this challenge. For information on payment policies, setting up your profile to receive payments, and general payment questions, please refer to [Payment Policies and Instructions](#).

### ELIGIBLE EVENTS:

[2021 Topcoder\(R\) Open](#)

### REVIEW STYLE:

#### Final Review:

Community Review Board ?

#### Approval:

User Sign-Off ?

### CHALLENGE LINKS:

[Review Scorecard](#) ?

### CHALLENGE TERMS:

[Standard Terms for Topcoder Competitions v2.2](#)

[Competition Non-Disclosure Agreement](#)

### SHARE:



## Recommended Active Challenges

[All Active Challenges](#)

Machine Learning   Data Science   Other



MM

**Rodeo II Sprint: Sub-Seasonal Climate Forecasting - temp56 Task, period #22**

Ends Sep 19

Prize Purse \$1,375

Submission

 38  16

1d 7h to register

Machine Learning

Data Science

Other

MM

**Rodeo II Sprint: Sub-Seasonal Climate Forecasting - temp34 Task, period #22**

Ends Sep 05

Prize Purse \$1,375

Submission

 39  16

1d 7h to register

Machine Learning

Data Science

Other


MM

**Rodeo II Sprint: Sub-Seasonal Climate Forecasting - prec56 Task, period #22**

Ends Sep 19

Prize Purse \$1,375

Submission

 51  17


1d 7h to register

## Recommended THRIVE Articles

[Explore THRIVE](#)


5 min

**List of awesome learning res..**

 Aug 03, 2020

5 min


**Is Data Science Just a Rebra..**

 Aug 03, 2020



5 min

## Basic Linear Model Fitting w..

 Aug 03, 2020



COMPETE  
TRACKS

COMMUNITY  
HELP CENTER

ABOUT



© 2020 Topcoder

[Policies](#)

