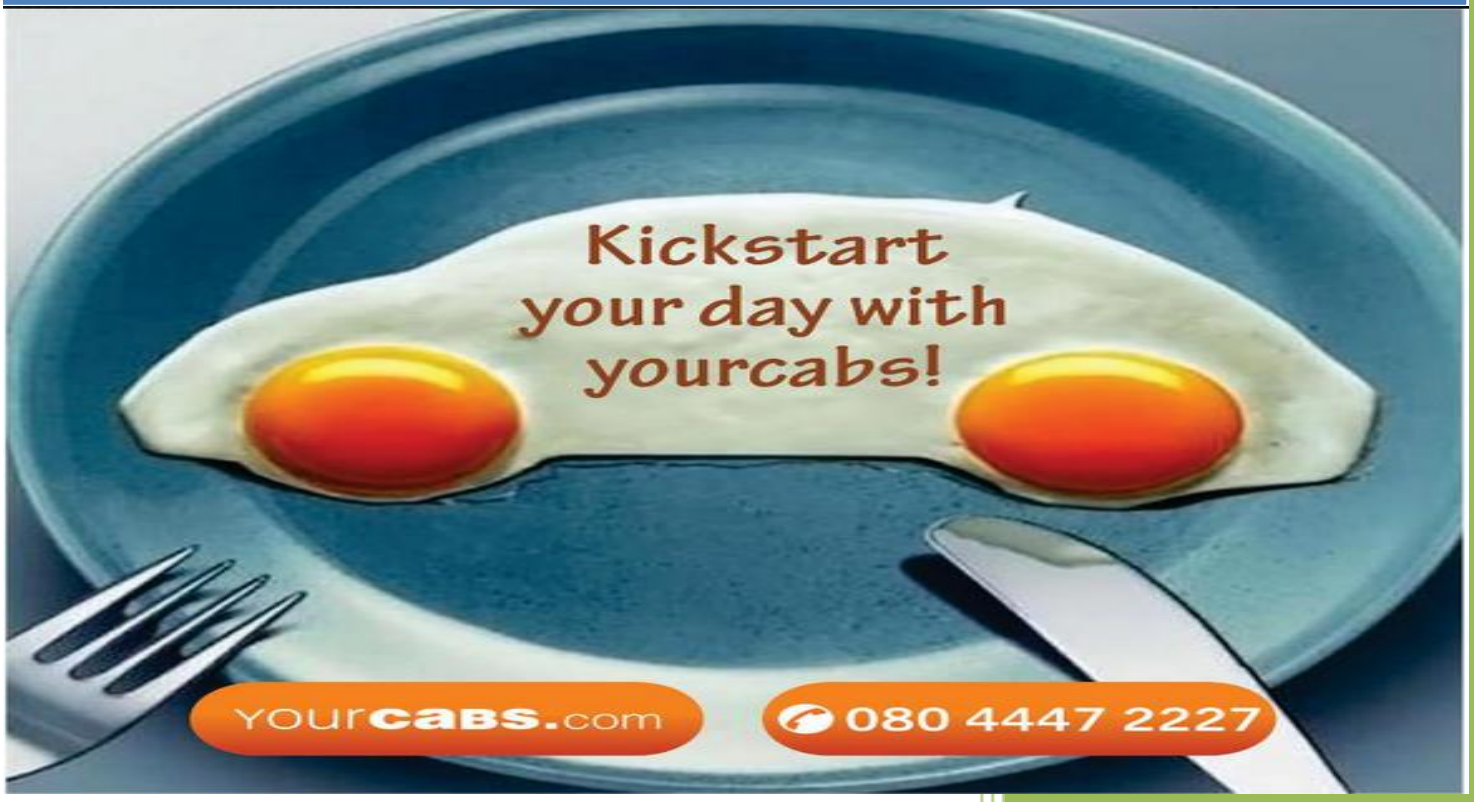
Planning for Future at Yourcabs



FCAS Group- B2

Deepti Singh	61410221
Nikhil Gupta	61410764
Piyush Gupta	61410280
Piyush Singh	61410800
Shriya Shekhar	61410004

Problem Description

Our client Yourcabs.com operates a platform to efficiently connect consumers in need of transport, with vendors in need of increased occupancy. Customers can make bookings through 3 possible methods- Online, Mobile website and phone. While traditionally most of the bookings have been coming from the phone method, the other 2 methods are fast catching up in terms of no of bookings being made. To maintain infrastructure to cater to a certain demand level in any of the booking methods, certain fixed and variable cost is incurred by our client. As such, matching demand with optimal infrastructure for each booking method is the first step in minimizing these costs. Various costs associated with different booking methods are given below:

Phone: Maintaining operators and phone lines. This is dictated by the client management agreed customer service levels for each demand point.

Online Booking: Dedicated servers needed to cater to the online traffic. A good forecast on an hourly basis can help us to maintain a base level of infrastructure and then rent optimal server capacity to cater to any predicted peaks.

Mobile Website Booking: Dedicated mobile servers to cater to the mobile traffic. As with online bookings, a good forecast can help to maintain a base level of infrastructure and then rent optimal server capacity to cater to any predicted peaks.

There are asymmetric costs associated with demand forecasts. Over forecast will lead to excess spending in capacity infrastructure while an under-forecast will lead to customer dissatisfaction, which might lead to lost sales in this competitive market. This loss is increased manifold if we look at the customer lifetime value (LTV).

This report aims at minimizing the tradeoff by maintaining optimal infrastructure for each method. This is done by forecasting future demand through each method by making use of available 3 years data.

We have worked on bookings made through the 3 modes in the past 3 years. This data was made available through the client booking systems. We have used time series which contain number of bookings made on an hourly basis by each mode. The following graphs show that the there has been an increasing trend in each of the booking methods. Although seasonality and level is

difficult to predict with just this data, we also see noise in terms of various external factors such as holidays, missing data at various time stamps.

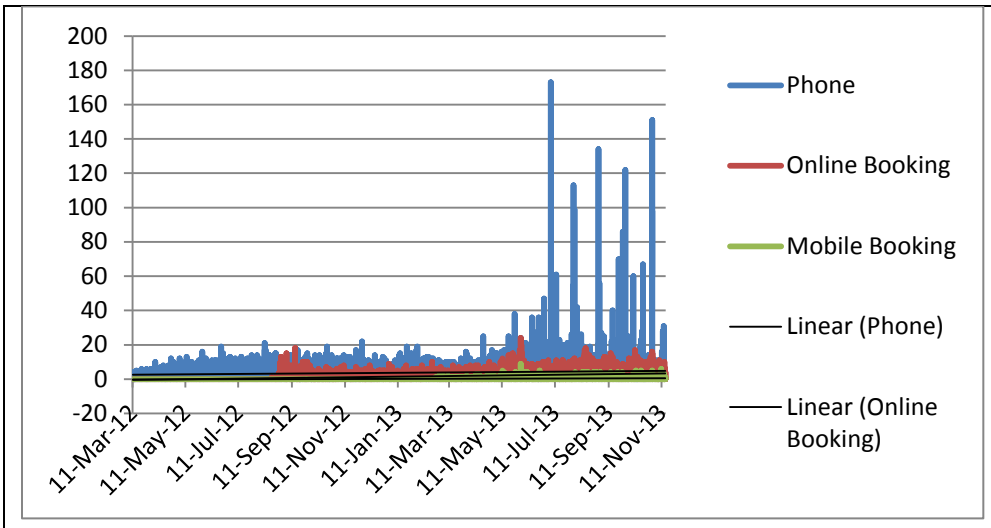


Fig 1 Trends of bookings through different methods

Figure 1

Owing to the large asymmetrical costs associated, we had to ensemble the data by using multiple regression as well as the Holt Winter's method for each series. We have taken weighted averages to finally reach a forecast value for the demand for each method. Performance of the model has been modeled based on the RMSE. We have further compared the predicted results with a naïve forecast to check if the predicted model was in fact needed.

In short, a model is proposed which can be used at various future times to predict the demand routed through each booking method. This information can be clubbed with the other market information and desired customer service levels for each booking method to maintain optimal infrastructure.

Methodology

This business problem aims at capturing future demand based on the existing data of 3 years. Although the data is available for 3 years, we had to make a choice on the best available model, based on the amount of data available for each method. Following report captures the data preparation, forecasting method used and performance evaluation methodologies in brief.

Data Preparation

Client provided us details of each of the booking which included the time stamp of booking, mode of booking along with route and cab details. It was assumed that the mode of booking

other than mobile and online was phone. The business problem required granularity at hourly level, so the data was aggregated accordingly. Till 10th March 2012, data was available only on daily basis hence this data was rejected. Moreover, the introduction of online booking in late August changed the pattern of phone booking which was the only mode of booking till then and hence the data from 1st September 2012 is used for forecasting for phone and online booking. For mobile bookings data after 1st February to forecast bookings as the mode was not introduced in February 2013

The time series plot of hourly bookings on each mode exhibited unusual spikes on holidays. The data for these holidays was removed and then replaced by the 4 day centered moving average for the same hour. The dates for which the data was smoothened were: 6th July 2013, 1st August 2013, 30th August 2013, 22nd August 2013, 27th August 2013, 30th September 2013, 9th October and 31st October.

Method

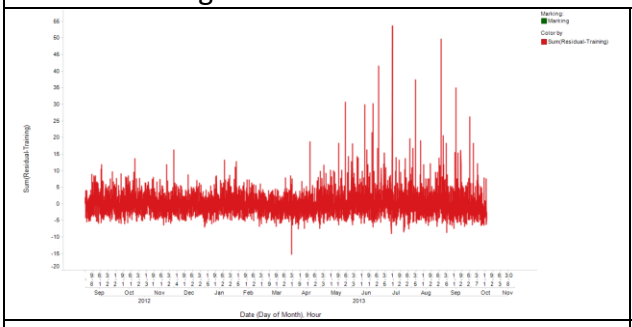
Visualization of the time-series plot (on hourly aggregated data) of phone booking suggested that data has hourly seasonality with slight upward trend. To confirm this, we de-trended the data using first difference method and found that residuals has hourly seasonality. The data was then partitioned with 720 points (30 days) in the validation period and rest of the 9864 data points in the training period. To capture the hourly seasonality, a multiple regression model was run to predict the number of bookings with date and 23 dummies for 24 hours as input variables and hourly number of bookings as output variables. The linear regression captured the long-run trend and seasonality.

To account for the dynamically changing trend and seasonality, Holt-Winters no trend model was fitted as the data did not show either additive or multiplicative seasonality. The models were then combined with weights computed by minimizing the error values. Thus an ensemble was created as a weighted average of Holt-Winters and Multiple Regression models. Details of the model used are given in the appendix.

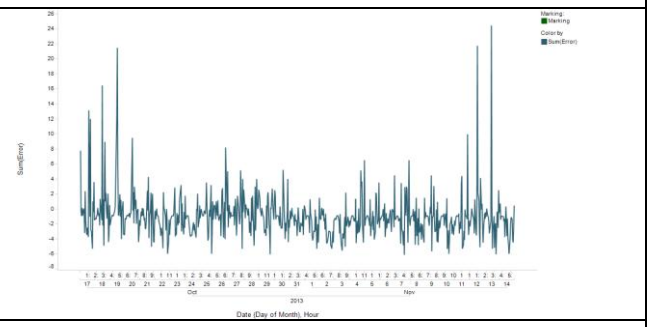
Model for online booking was also created using the same approach whereas for mobile booking, only Holt Winter's method with no trend was used because its performance on both training and validation data for mobile booking was satisfactory.

Model Evaluation: Residuals from each of the model was plotted against time. There was no observable trend in residuals. Also, we plotted the ACF for each residual and found that there was no significant correlation among residuals.

Phone Booking



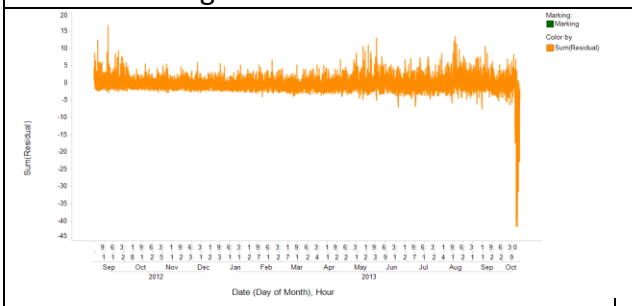
Training Data



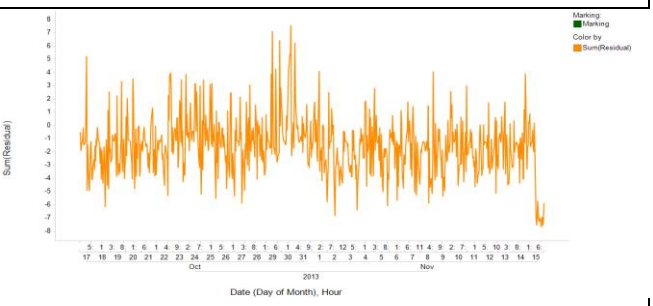
Validation Data

Figure 2

Online booking



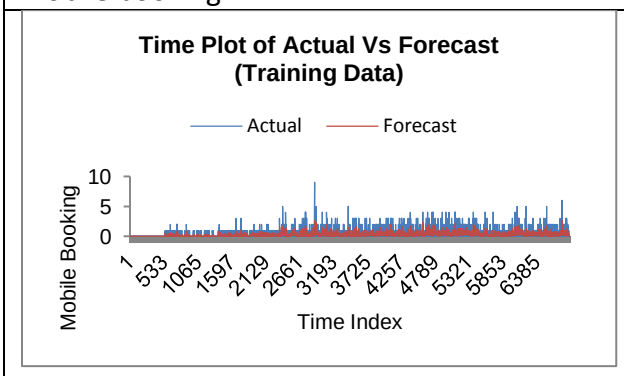
Training Data



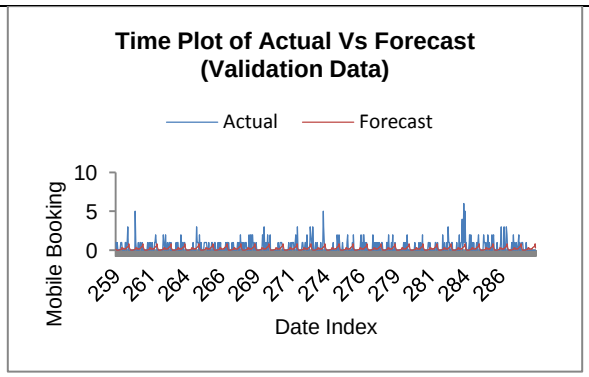
Validation Data

Figure 3

Mobile booking



Training Data



Validation Data

Figure 4

Benchmarking

We benchmarked our model against Naïve forecast and observed that the errors produced by the model were less as compared to the Naïve forecasts. RMSE was chosen to benchmark the model

because it is robust with zero values and with the data having large variations. Our booking data has both of these issues. The following table compares the RMSE of the model with the naïve forecast.

	Phone Booking	Online Booking	Mobile Booking
Training	2.99	2.65	5.24
Validation	3.17	2.68	0.78
Naïve(Validation)	4.18	3.08	0.98

Table 1: RMSE comparison with Naive forecast

We also compared the forecast of our model with the naïve forecast and the actual data in the validation period. The comparison of these forecasts is shown in appendix. Visual observation of these forecasts suggests that the developed model performs better than naïve forecasts. Both the developed model and the naïve forecasts do not predict the peaks accurately. Since this model would be used to create the base infrastructure (on an hourly basis) for the next 1 month, we believe that using this model would be appropriate.

Limitations

While we have tried to forecast for the usual days, our model is limited in terms of the following cases:

- Not a good predictor for holiday demand
- Extreme values cannot be predicted by the model. However, the model is robust for predicting usual demand.

Appendix

Model Details

PHONE BOOKING		
Multiple Linear Regression		
Output Variable		
Predictor Variable	Continuous	Date-index (to capture linear trend)
	Dummy	Hour_0, Hour_1, …., Hour_22 (23 dummies to capture hourly seasonality)
$F_{t+k} = \alpha + \beta(\text{Date\_index}) + \beta_0(\text{Hour\_0}) + \beta_1(\text{Hour\_1}) + \beta_2(\text{Hour\_2}) + \ldots + \beta_{22}(\text{Hour\_22})$		
Holt's Winter exponential smoothening		
$F_{t+k} = (L_t) \times \text{Season(intra-day)}_{t+k-24}$ • $L_t = \alpha Y_t/[\text{Season(intra-day)}_{t-24}] + (1-\alpha)(L_{t-1})$ • $\text{Season(intra-day)}_t = \Upsilon[y_t/L_t] + (1-\Upsilon) \times \text{Season(intra-day)}_{t-24}$ • $\alpha = 0.2 \quad \Upsilon = 0.05$		
Ensemble		
$F_{t+k} = 0.9\times(F_{t+k})^{MLR} + 0.1\times(F_{t+k})^{\text{Holt's-Winter}}$		

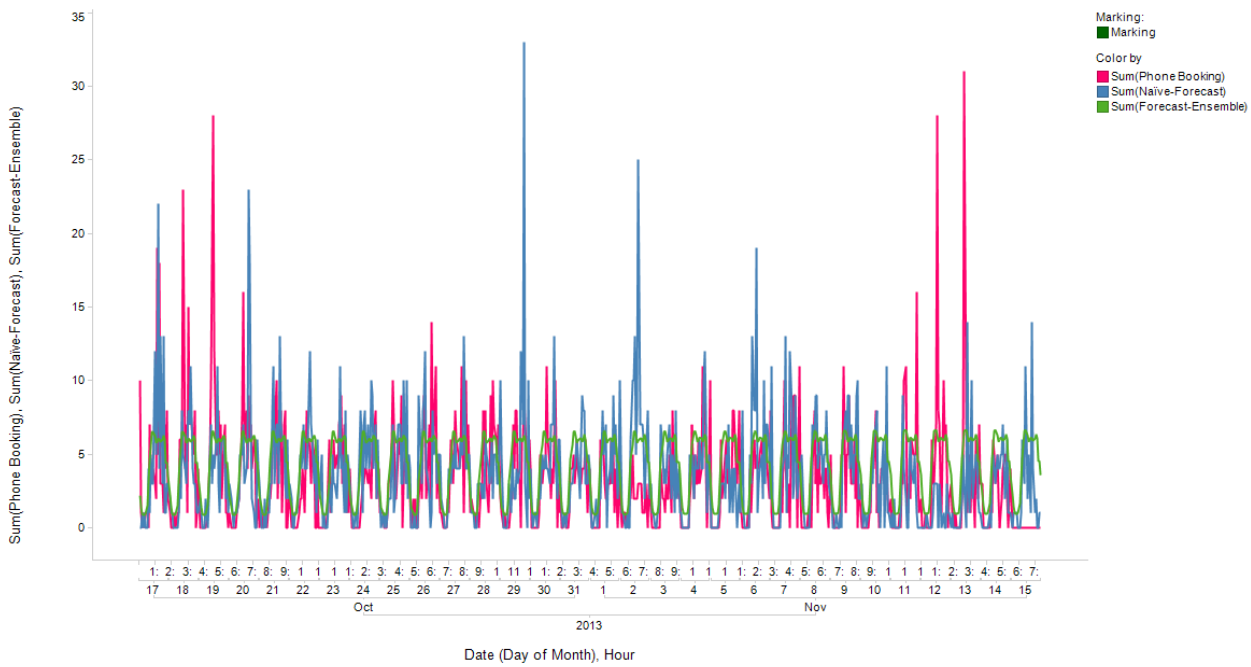
ONLINE BOOKING		
Multiple Linear Regression		
Output Variable		
Predictor Variable	Continuous	Date-index (to capture linear trend)
	Dummy	Hour_0, Hour_1, …., Hour_22 (23 dummies to capture hourly seasonality)
$F_{t+k} = \alpha + \beta(\text{Date\_index}) + \beta_0(\text{Hour\_0}) + \beta_1(\text{Hour\_1}) + \beta_2(\text{Hour\_2}) + \ldots + \beta_{22}(\text{Hour\_22})$		
Holt's Winter exponential smoothening		
$F_{t+k} = (L_t) \times \text{Season(intra-day)}_{t+k-24}$ • $L_t = \alpha Y_t/[\text{Season(intra-day)}_{t-24}] + (1-\alpha)(L_{t-1})$ • $\text{Season(intra-day)}_t = \Upsilon[y_t/L_t] + (1-\Upsilon) \times \text{Season(intra-day)}_{t-24}$ • $\alpha = 0.2 \quad \Upsilon = 0.05$		
Ensemble		
$F_{t+k} = 0.8\times(F_{t+k})^{MLR} + 0.2\times(F_{t+k})^{\text{Holt's-Winter}}$		

MOBILE BOOKING
Holt's Winter exponential smoothening

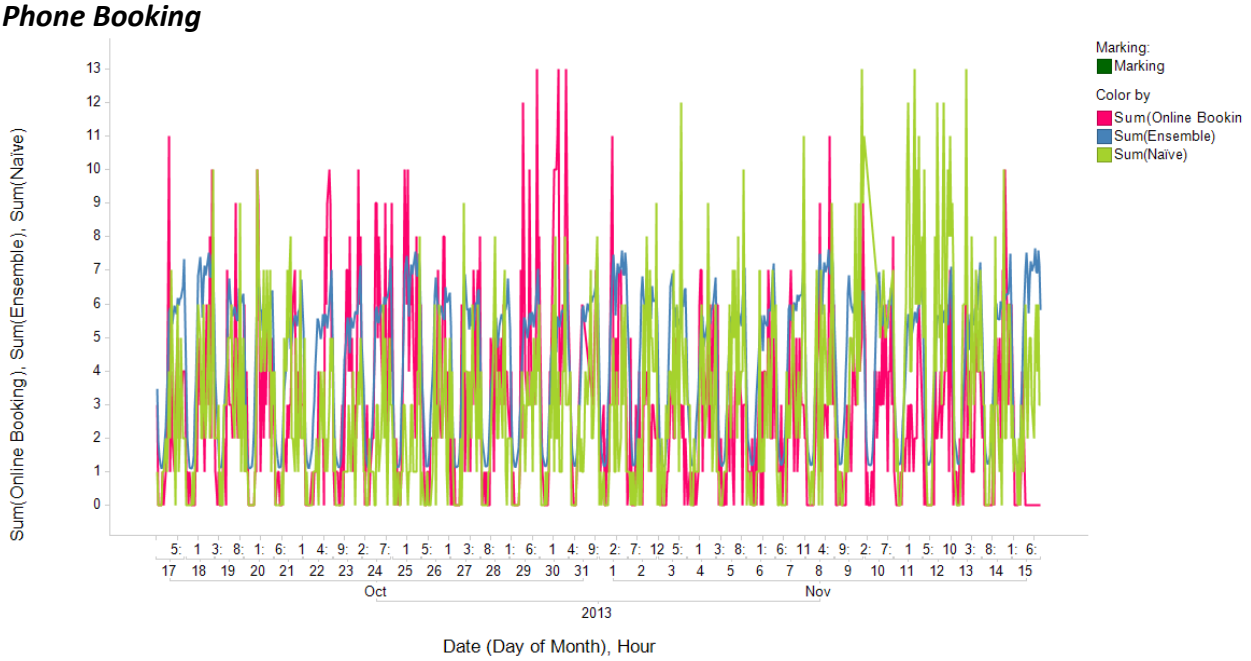
$F_{t+k} = (L_t) \times \text{Season(intra-day)}_{t+k-24}$

- $L_t = \alpha Y_t / [\text{Season(intra-day)}_{t-24}] + (1-\alpha)(L_{t-1})$
- $\text{Season(intra-day)}_t = \Upsilon[y_t/L_t] + (1-\Upsilon) \times \text{Season(intra-day)}_{t-24}$
- $\alpha = 0.2 \quad \Upsilon = 0.05$

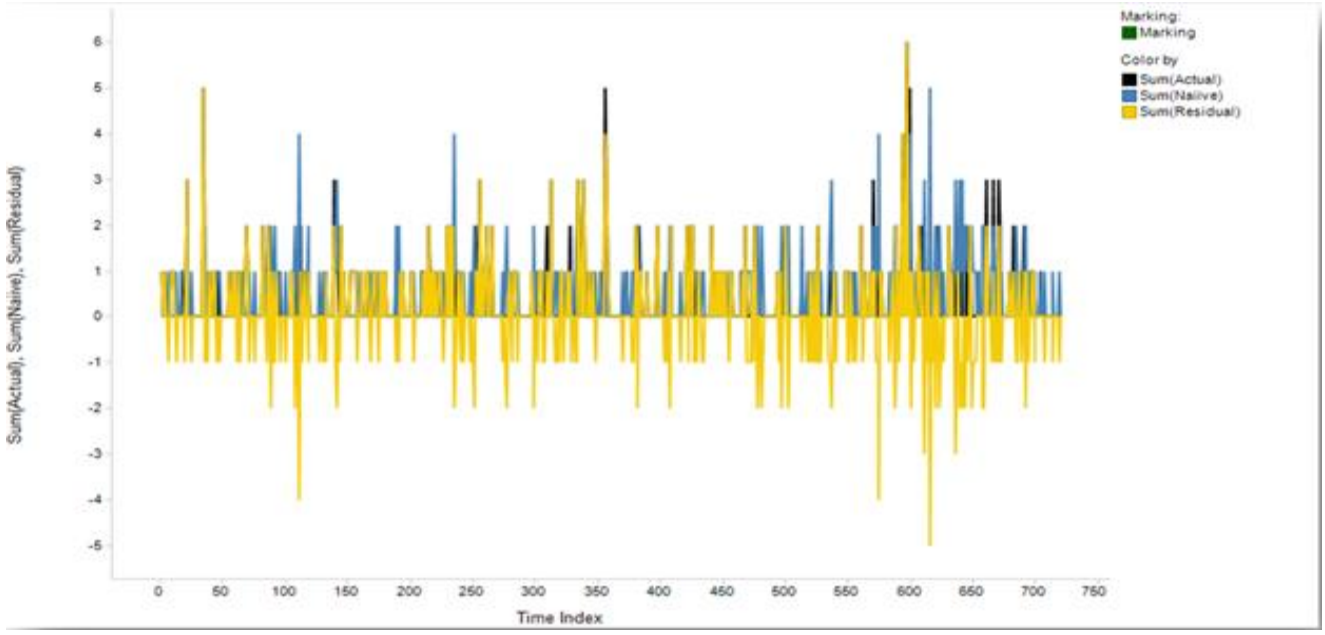
Forecast over validation



Phone Booking



Online Booking



Mobile Booking