2013

Identifying individual usage patterns for an effective promotion strategy

yourcabs.com

Group-B1
Indian School of business
12/25/2013

Executive summary

The report describes an analytics approach towards designing better promotions for increasing the revenues of yourcabs.com. The two key questions answered by this study are

1. When to launch a promotional campaign
2. Which customers to target to have the maximum ROI on the promotional campaign

The answer the above question two set of forecasting models have been built

1. A linear regression model to forecast weekly revenues.
2. A logistic regression model to forecast the chance of making a booking the next week.

The data set available does not contain revenue per transaction hence revenues are estimated basis the time of travel calculated by the journey start and end time. Due to unavailability of quality data we restrict our study to the time period July 2013 to November 2013. Also, for the purpose of the report we restrict our analysis for the point to point business model.

The data series used for developing the forecasting model are

1. From and to dates
2. Number of booking per user

Forecasting revenues: The plot of the daily revenues shows no trend but a daily seasonality in the data. Also since the data small we use regression model based forecasting. On analyzing multiple models the best predictors for revenue emerged to be the last seven days revenues. Using these predictors we developed a regression model for predicting the revenue. These revenue forecasts can be used as indicators for comparing the revenue targets and the forecasts. In case of a falling revenue forecast appropriate promotional campaigns can be launched to increase bookings.

Forecasting usage pattern: In order to identify the usage pattern for each customer two approaches have been tried.

1. Developing a logistic regression model for each user: All repeat customers can be segregated in the data and for each user a simple logistic regression model is developed taking the lags of number of bookings in the last week as the predictors. This model generates a binary output signifying whether the said user will book a cab the next week.
2. Developing a logistic regression model for all repeat customers: The simple model developed for each user can be extended by using interaction variables for each user to develop a common model that can be used for predicting the possibility of making a booking for each user.

The above two analytic models will help YourCabs to decide the correct time to launch a promotional campaign and also help in correctly targeting the campaign to the right customers to increase customer loyalty and increase repeat usage for your cabs services.

Introduction

YouCabs.Inc is a taxi hire company operating out of Bangalore. That operates on a market place model and integrates multiple taxi service providers. The company works on an interesting model, wherein your cabs oesnot own cabs but have short term arrangements with taxi owners to providing cab service. The dataset received from YouCabs.Inc. contains time series information on boking id, vehicle model type, type of travel package, type of travel, unique identifier area, unique identifier of city, type of booking (desktop/mobile), booking status, cancellation purpose, date/time of booking, longitude/latitude of area. We adopt a time series forecasting approach to develop analytical models for solving the following business problems.

Business Problems

The two key business problems addressed in the report are

1. When to launch a promotional campaign
2. Which customers to target to have the maximum ROI on the promotional campaign

For deciding when to launch a promotional campaign we forecasted revenue on a weekly basis .These forecasted values form the basic indicator for comparing the future revenues with the internal target of your cabs. For forecasting revenues we use the time of travelled as the proxy for distance. Distance travelled is calculated by considering an average of 30Km per hour for each point to point booking. To calculate the number of weekly bookings by each customer we extracted information for each user and aggregated the usage for each week. The resulting dataset was studied for the usage pattern of different customers & tried to predict their future usage probability. By identifying the better prospective customers & their usage pattern, we can customize our offers for individual customers so as to increase their repeat usage of cabs from our company. It will also help us optimize our marketing & promotional expenditures.

Forecasting Problem

To solve the above mentioned business problems we develop two forecasting models

1. A linear regression model to forecast weekly revenues.
2. A logistic regression model to forecast the chance of a particular user making a booking the next week

The key challenge in solving the above two problems is that we do not have data for revenues. To address this we estimated the revenues using the time of travel as the proxy for distance travelled. Distance travelled is calculated by considering an average of 30Km per hour for each point to point booking. The resulting data set was used for forecasting revenues.

To calculate the number of weekly bookings by each customer we extracted information for each user and aggregated the usage for each week. The resulting data set contained usage of each user aggregated over week duration. This data set was used for developing the usage forecasting model.

Data visualization

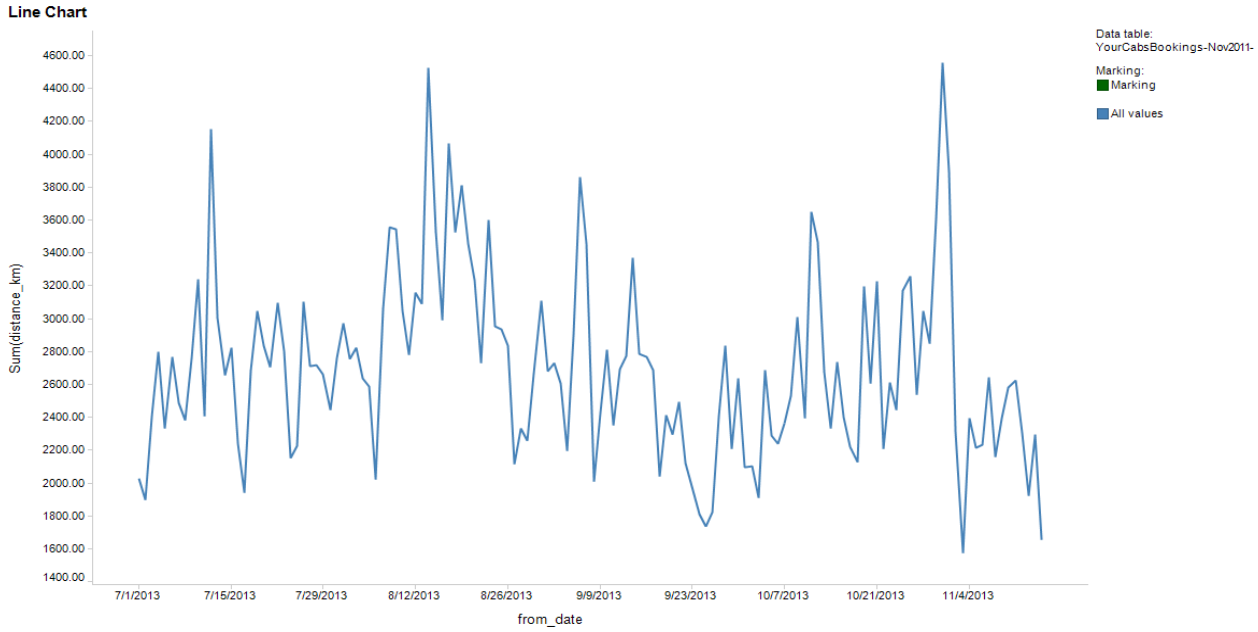

Fig 1:Plot of sum of distance vs day

In the graph, along Y-axis,we have plotted distance (a proxy) & X-axis is from date. From the data plotted above, there is apparently no trend, but the plot shows some seasonality, to understand the seasonality a week of the day plot for the day showed a seasonal pattern every week

Data preparation

The complete dataset contains booking data from 2012 onwards. For the purpose of calculating revenues we require to and from date and time. This information is available only for juy 2013 to November 2013. Hence this is the time period considered for this study.

The data set is further filtered for only point to point service and removing any incorrect data.

The resulting data set contained 20000 bookings to be analyzed.

Developing model for forecasting revenues

The plot for revenues does not show any trend and after attempting multiple forecasting methods a multiple regression model emerged as the best suited model for this purpose. The predictors that generated the best model were lag1 to lag7 for the number of booking on each day. This covered for the daily seasonality observed in the data set.

The resulting model is as follows

Forecasting model

$$Y = \beta_0 + \beta_2 * W_{t-1} + \beta 3 * W_{t-2} + \beta 4 * W_{t-2} \ldots\ldots \beta 5 * W_{t-7}$$

Evaluating the forecasted model :

The errors summary for the training and the validation periods are as follows

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
45727135575	20388.7526	-0.00153727

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
45706605278	46653.02791	-9566.88492

The resulting forecast is benchmarked against the naïve forecast. The plot below show the forecasted, actual and the naïve forecast of the complete data set. Also the residuals are plotted and checked for autocorrelation. The residual plot does not show any remaining seasonality or trend in the data.

From the model above, we observe that naïve is a better predictor than the developed forecasting model.Our developed forecasting model although fits well but still does not capture the extreme peaks & valleys well.

Developing forecasting model for customer usage:

We used logistic regression for forecasting customer usage as it involved a binary decision of whether a particular customer will be booking a cab in a future week or not**.** We have not used

neural network method as number of data points for each customer is quite less & the method would not have yielded a robust result. For the purpose of this, a simple logistic mode is used with the number of booking by the a particular user in the previous week and the lag1 of this predictor. The resulting simple logistic model looks like

Logistic regression model :

Logit(week=1) = $\beta_0 + \beta_1 * W_{t-1} + \beta_2 * W_{t-2}$

The resulting coefficients for the model are as follows

The cutoff probability for this case is taken as 0.5

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	13.96868706	1079.650391	0.98967713	*
LAg1	57.60430527	2256.641602	0.97963494	1.04048E+25
Lag2	0.84831196	1358.720825	0.99950182	2.33570075

Evaluating the forecasted model :

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	1	2
0	2	1

Error Report			
Class	# Cases	# Errors	% Error
1	3	2	66.67
0	3	2	66.67
Overall	6	4	66.67

Fig 4 :confusion matrix & error report for logistic regression (individual customer)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	3	0
0	3	0

Error Report			
Class	# Cases	# Errors	% Error
1	3	0	0
0	3	3	100
Overall	6	3	50

Fig 5 :confusion matrix & error report for naïve forecast (individual customer)

From the above report,we find that naïve is a better predictor as % error is less than the output from logistic regression.But ,then we need to take an economic decision in view of our business objective i.e develop customized marketing pitch for individual customers to enhance his number of bookings.As we see,in case of naïve forecast we need to target 6 customers whereas in case of logistic regression,it is 3 customers only.So ,clearly in case of logistic regression ,our expenditure on marketing & sales is low in abosolute terms.So from business point of view,logistic regression gives a better model.

Developing forecasting model for multiple customer usage:

For predicting the usage pattern for multiple users,we used the same lag 1 & lag 2 for number of weekly bookings for user & used variables (D1,D2 , ,Dn) for users

to create interaction terms as we observed different trend for different users.

The forecasted model :

Logit(week=1) = $\beta_0 + \beta_1 * D_1 * W_{t-1} + \beta_2 * D_1 * W_{t-2} + \beta_3 * D_2 * W_{t-1}$ ..... $\beta_{2n} * D_n * W_{t-1}$

For the purpose of this model we are using a cut-off probability of 0.5 for predicting the binary outcome of a cab booking by a user.

Evaluating the forecasted model :

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	11	8
0	0	8

Error Report			
Class	# Cases	# Errors	% Error
1	19	8	42.11
0	8	0	0.00
Overall	27	8	29.63

Fig 6 :confusion matrix & error report for Logistic regression (multiple customers)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	6	4
0	13	4

Error Report			
Class	# Cases	# Errors	% Error
1	10	4	40.00
0	17	13	76.47
Overall	27	17	62.96

Fig 7: confusion matrix & error report for naïve forecast (multiple customers)

From the above we find that logistic regression is fairly close to naïve in terms of accuracy & also accuracy level has increased vis-à-vis model for single customer.

The above model thus generates one easy to implement model that can be used for forecasting the usage pattern for each user.

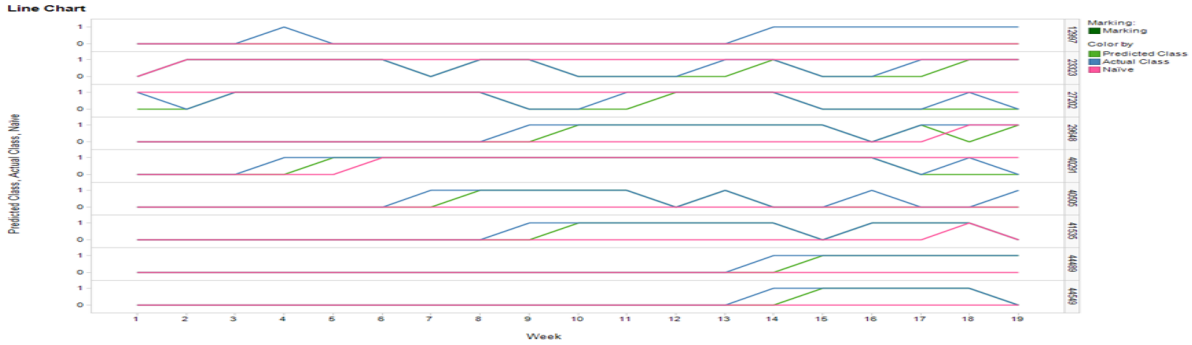

Fig 8 :Plot of actual,predicted & naïve for multiple customers

Recommendations:

Using the time series forecasting techniques we have developed two model the recommendations from the resuts of each of the model is as follows

Revenue forecast: The resuling regression model predicts the future weekly revenues. This forecast can be used as a indicator for comparing the internal target and performance of your cab. Should the forecasted revenue fall below the internal target it is suggested to launch a promotional campaign such as discount coupons, targeted advertising, calling customers ir sending promotional emails.

Usage forecast: The model for forecasting usage pattern is to be used predict the probability of a user booking a cab in the next week. This probability presents the guideline for targeting customers for the promotional campaigns. This method needs to be tested to find the probability range that generates the highest returns for the promotional campaign. For the test purposes targeting the campaigns for users having a usage probability of 0.6 to 0.8 is recommended.

The model can be automated and will generate more accurate forecasts as the quality and quantity of data improves.

Appendix

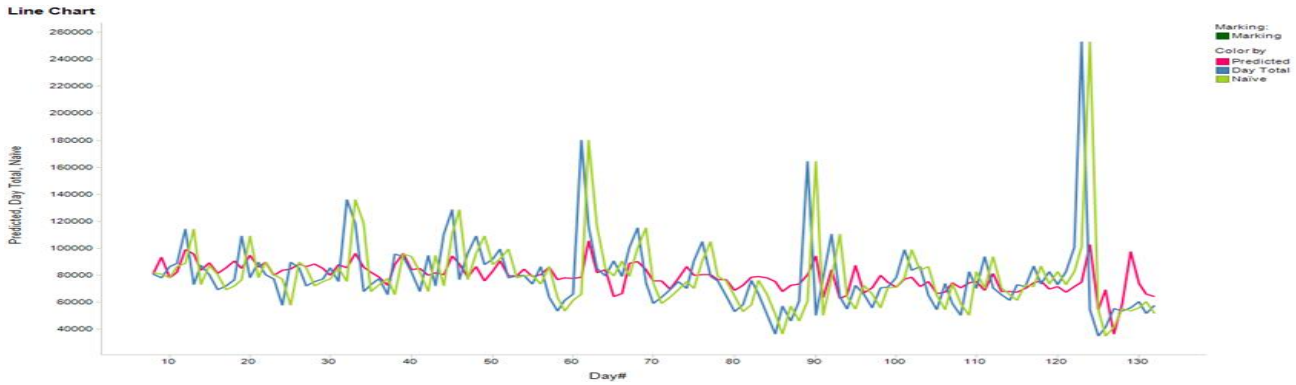

Fig 2.Plot of actual,naïve & predicted revenue

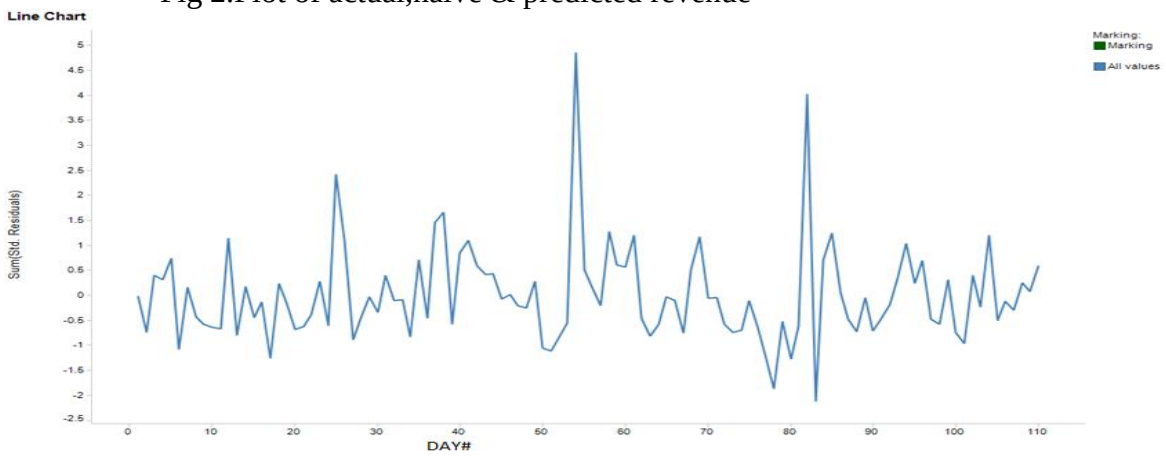

Fig 3: Plot of residuals of revenue

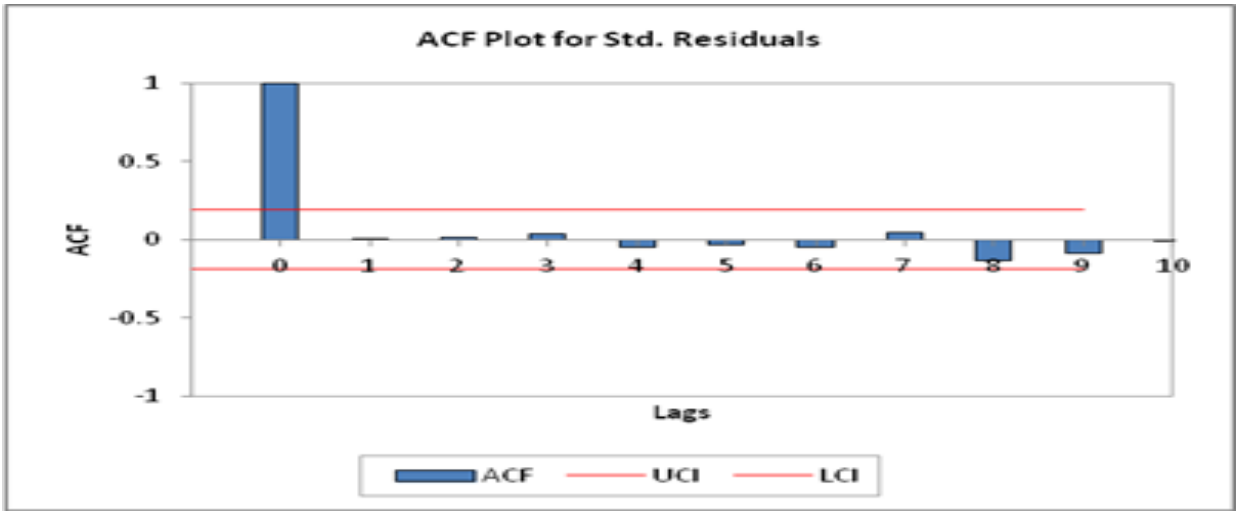

Fig 9 : ACF plot for residuals for revenue forecast

Revenue model

Values for each coefficient for revenue model are as follows

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	49866.98828	16484.36719	0.00314527	7.07618E+11
Lag 1	0.23313747	0.09877834	0.02016694	2521144000
Lag 2	-0.02588247	0.09453377	0.78479916	56481210
Lag 3	0.10206722	0.09340308	0.27707306	477876400
Lag 4	-0.12229206	0.09344744	0.19358641	862063600
Lag 5	-0.03870578	0.09368011	0.68034959	196355.5313
Lag 6	0.17851418	0.09383794	0.05994384	1883915000
Lag 7	0.05024991	0.09343741	0.591892	129659000

Combined Usage model for each user

The resulting coefficients for the above model are as follows

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-0.66924739	0.77623022	0.38859043	*
D1Count_Lag1	1.06582153	156.1972961	0.99455565	2.90322304
D2Count_Lag1	1.42652535	524.6663818	0.99783063	4.16420507
D3Count_Lag1	17.00691986	1596.084229	0.99149835	24322660
D4Count_Lag1	17.01086044	1756.925293	0.99227488	24418710
D5Count_Lag1	1.99638903	731.057312	0.99782109	7.36242294
D6Count_Lag1	18.74422836	3279.363281	0.99543947	138202200
D7Count_Lag1	9.27117729	1700.464966	0.99564981	10627.25098
D8Count_Lag1	9.55587292	2555.291992	0.99701619	14127.41699
D1Count_Lag2	1.15863025	194.2185974	0.99524015	3.1855669
D2Count_Lag2	0.35530999	849.3411865	0.99966621	1.42662287
D3Count_Lag2	0.18747158	1390.560303	0.99989241	1.20619595
D4Count_Lag2	8.93427277	1348.86084	0.99471521	7587.615723
D5Count_Lag2	0.47059011	1196.871216	0.9996863	1.60093868
D6Count_Lag2	18.53421593	3616.025391	0.99591041	112023100
D7Count_Lag2	17.16884422	2420.070557	0.99433959	28597930
D8Count_Lag2	18.29940605	3404.552002	0.99571139	88579080