2013

Predicting Customer Cancellation of Cab Bookings for YourCabs.com

BADM Final Project

Group A9 – Big Data

Anupama Atmuri	61410226
Garrett Butler	61419017
Leena Bhai	61410035
Priyanka Paul	61410093
Rohith Lokareddy	61410872
Shweta Agarwal	61410422

Indian School of Business
12/1/2013

Executive Summary

Our Client, YourCabs is a Bangalore-based technology platform that aggregates fleet owners and vehicles, in the car rental space. The company, founded by Rajath Kedilaya in 2011 has managed to create an intelligent network that manages real-time supply and demand of cabs. In this assignment we will try to predict possible cancellations of cab booking by the customer using data obtained from the company. Our goal is to reduce the cost incurred by the company as a result of cab cancellations made by the customer. By predicting possible cancellations an hour before the pickup time, YourCabs will be better able to manage its vendors and drivers by providing them with up to date information about customer cancellations and reduce the cost incurred from sending a cab to a booking location that has been cancelled by the customer.

Accurate prediction of customer cancellations will lead to a reduction in company costs. If we assume that the cost of sending a cab for a booking that will be cancelled by the customer is Rs 100, and the cost of calling a customer flagged by our model an hour before the pickup time to confirm the booking is Rs 10. For each possible cancellation that is predicted accurately, YourCabs will save Rs 90. If the model incorrectly predicts a customer cancellation, it will cost YourCabs Rs 10 to call the customer to confirm the booking. Success would be defined as a reduction in overall cost to the company for cab cancellations from the customer end.

Our data analysis model used several methods to analyze the data including classification tree, K-nearest neighbor, Naïve Bayes and Ensemble. The accuracy of the model coupled with the final business goal of reducing cost for the company was used to finalize the model for the prediction. The model that we selected in the end was Naïve Bayes. Not only does the model have an overall low error rate, but also the cost incurred by the company using this model is the lowest.

Our recommendation includes running the model in real time on an hourly basis for all pickup times, which are within an hour's time. The model will flag all likely booking cancellations and the operator will call the customers to confirm the booking. Once the operator receives confirmation from the customer, the cab will be dispatched to the pickup location. By using the model for predicting possible customer cancellations, the company will successfully reduce the cost incurred from sending a cab to a pickup location where the customer is not present.

Problem description

Every year the Company, YourCabs loses money due to customer cancellations. The company currently does not have a mechanism to track or predict these cancellations. The company currently only realizes that there is a cancellation when the cab reaches the location; resulting in a cost which can be quantified in such metrics as fuel cost, driver's salary, cab utilization, lost time that the driver which could have been spent attending other bookings and most important lower utilization by the vendors using YourCabs service. This also increases the variable waiting time by the customer. The cost of the cancellations due by customers on a yearly basis is calculated by assuming the average cost of

cancellation being Rs. 100 and on average 10% of all booking are cancelled by the customer. This equates to roughly 4,35,000 (43,50,000 bookings * 10%*100) in cost each year.

Our model is meant to predict if the customer will be cancelling his or her booking by not being at the specific location at pickup time. Changing the cost of calling a customer and a customer cancellation affects our model selection and the number of customers flagged by our model. Under the current costs structure, the model we use a naïve Bayes, which results in YourCabs calling 1124 customers out of a test set of 13,336 records. Of the 1124 customers flagged as possible cancels, 252 customers will actually cancel. This results in YourCabs saving Rs13.950 in this data set, and Rs. 45,400 on an annual basis. Thus we are reducing the cost of customer cancellation on an yearly basis from 4,35,000 to 3,89,600. Higher savings could be achieved if cost of calling a customer is reduce from Rs. 20 to lower levels. This could easily be accomplished with the use of SMS confirmations, which could be automated by a computer program.

Data description

We used all the records from Kaggle_YourCabs_Training_With_Cancellation_Reasons.csv (43432 records). Details about each of the input and output variables are shown in Exhibit 1. A sample dataset containing the input columns for the model is shown in Exhibit 2.

Data preparation

Our business goal was to be able to reduce the cost to company due to customer cancellations. So we just preserved the data that helped us reach the goal.

Data preparation steps –

1. Removed – package_id, to_area_id, from_city_id, to_city_id, to_date, book_status, booking_closed, fom_lat, from_long, to_lat, to_long columns from the data as they were not relevant for analyzing and reaching our business goal.
2. We considered using vehicle_model_id to see if there are any relationships between type of vehicle and number of cancellations. After creating cross-tables in Spotfire, we realized that 75% of the bookings were for Tata Indica and 75% of cancelations are also from Tata Indica. We checked other vehicles and this type of relationship existed throughout the data set. Therefore, vehicle_model_id was not an important input variable.
3. We considered from_area to see if there is any relationship between number of cancelations and from_area. We realized that most cancellations were from airport and most of the bookings were also from the airport. We checked from_area and this type of relationship existed throughout the data set. Therefore, it is not an important input variable.
4. After visualizing the data in Spotfire and analyzing the data, we realized that 18% of the total cancellation came from 20% of the car booking pick up times which took place 24 hours after the time it was created. Your cabs data had instances when cars were booked 3-4 weeks ahead of

pickup time. Thus, we created a binary variable derived column - Is_Booking_Within_a_day to bin the cab bookings 24 hours in advance of the pickup time.

5. After analyzing the data in Spotfire, we also realized that cancellations were prevalent in certain months and in mid-day rather than in late nights or early mornings. Therefore, we created the derived columns - Binned_Pickup_Month and Binned_Booking_Month, Binned_Pickup_Time_Of_Day and Binned_Booking_Time_Of_The_Day and used these derived columns as input variables.
6. Our data mining goal was to build an algorithm that effectively predicts customer-initiated cancelation. To address this goal, we created a binary variable - Is_Customer_Cancelled, 1 = customer cancellation and 0 = no customer cancelation.

Data mining solution:

Since this was a classification problem we looked at Classification trees, KNN and Naïve Bayes. We did not use prediction algorithms such as Multiple Linear Regression or Exponential Smoothing.

To ascertain whether each variable would be used in our final model, we applied the following methods:

a. Ran classification tree to find the appropriate predictors with all predictors as inputs and Is_Customer_Cancelled as output.
b. Ran the KNN Algorithm with the predictors as input and Is_Customer_Cancelled as output.
c. Ran the Naïve Bayes Algorithm with the predictors as input and Is_Customer_Cancelled as output.
d. Ran an emsemble on these two algorithm output on the basis of probability.

The detailed error summary for all the three algorithms is shown below:

Emsemble	Naïve Bayes	KNN Algo
82.32%	81.59%	74.65%
8.17%	7.30%	15%
15.79%	14.94%	21.13%

The Cost calculations for all three methods is as follows:

	KNN	Naïve Bayes	Emsemble
Saving to the company	34700	25200	24200
Cost to the company	21400	11500	12190
Net Savings to the company	13300	13700	12010

We chose the KNN approach as it gave the maximum savings given our assumptions.

The Naïve Rule Cost - 43500* 0.10 * 100 = Rs. 4,35,000.

Implementation savings after we extrapolate the savings to all the data = 14000*3.33 = Rs. 45K per year.

Conclusions

Advantages – Our model uses existing infrastructure, under utilized employees and existing fixed costs. Implementation of our model into the daily operations would be simple and require minimal amounts of training. The results could be easily confirmed via data and the cost of trial would be minimal and have no negative effects on the company or for the customer. In addition, we believe this model could be further improved and the savings would increase in line with natural growth of the company.

Limitations – Effectiveness of the model strongly relies on the call center employees, implementation would require YourCabs to slightly alter the call center employees current behavior. Collection and upload of the data into our model would also have to be on a continuous basis, which might also require change within the company if this is not the current practice.

Overall – We feel that the result is indicative of a real life gains/savings that would occur when running such an exercise. Though the overall savings would not numerically register on a monthly balance sheet, if similar gains/savings occurred on a weekly or monthly basis, the total value for such endeavors would make a huge difference over a years time span. Therefore we feel this is a success and will use this as a positive lesson moving forward in the data mining realm.

Operational recommendations

- Run the model at one hour intervals to benefit the most from the model
- Send SMS to the driver confirming the customers booking with location
- Send SMS to the customers prior to the pickup time with vehicle details

Appendix

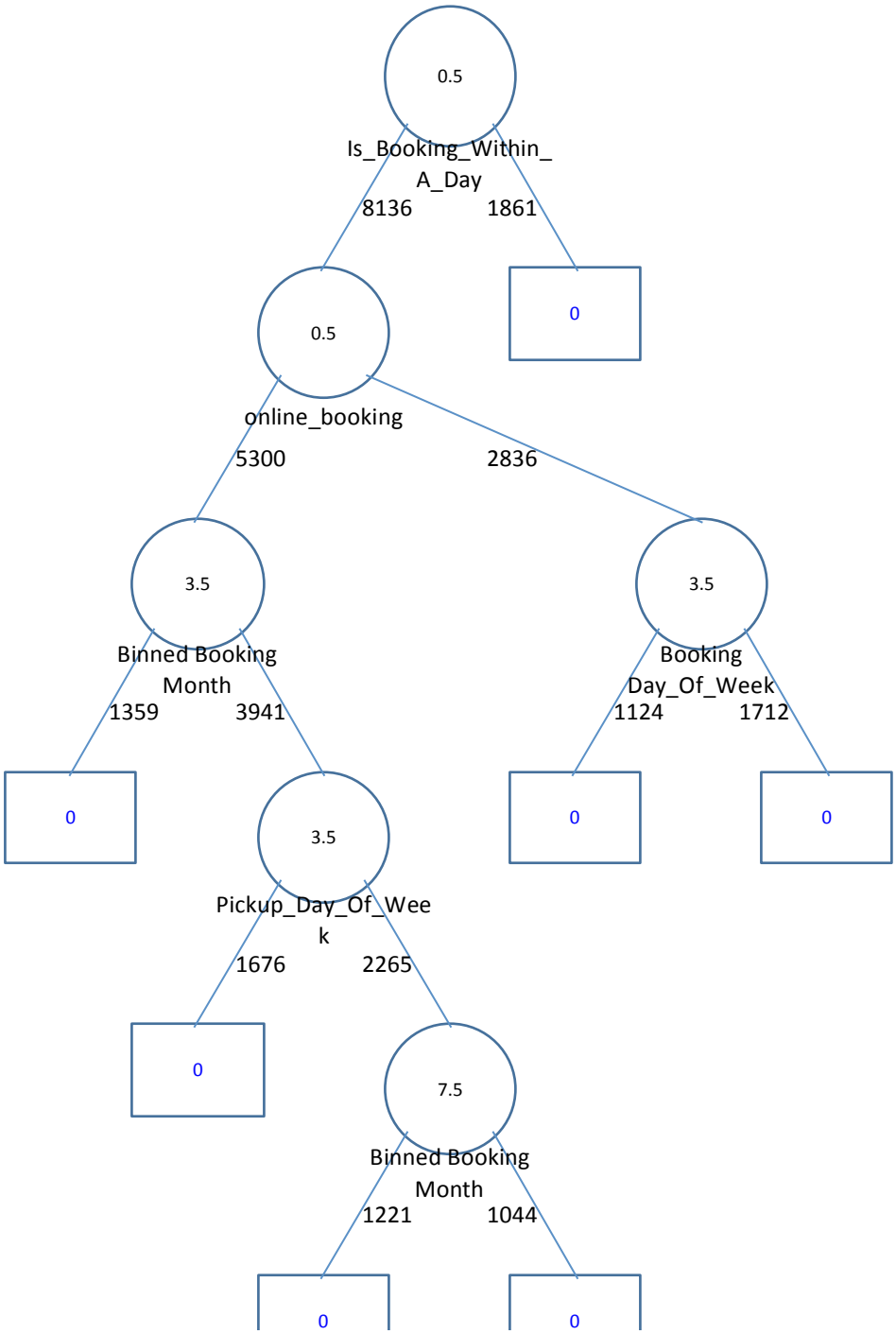
Exhibit 1: Data Description

Record - Each record carries information about a single cab booking at Your Cabs		
Output Variable		
Name	**Description**	**Value**
Is_Customer_Cancelled	Categorical variable; Derived from 'Cancellation Purpose' column	value = 1 => cancellation is done by the customer value = 0 => cancellation is not done by the customer
Input Variables		
Name	**Description**	**Values**
Booking_Day_Of_The_Week	Categorical variable; Derived from 'booking_created' column	value = 1-7 for Sunday to Saturday
Pickup_Day_Of_The_Week	Categorical variable; Derived from 'booking_created' column	value = 1-7 for Sunday to Saturday
Binned_Booking_Time_Of_The_Day	Categorical variable; Derived from 'booking_created' column	value = 1(12am – 4am) and so on upto value = 6(8pm – 12am)
Binned_Pickup_Time_Of_Day	Categorical variable; Derived from 'from_date' column	value = 1(12am – 4am) and so on upto value = 6(8pm – 12am)
Binned_Pickup_Month	Categorical variable; Derived from 'booking_created' column	value = 1 to 12 for January to December.
Binned_Booking_Month	Categorical variable; Derived from 'from_date' column	value = 1 to 12 for January to December.
Is_Booking_Within_a_day	categorical variable; derived column from 'booking_created' and 'from_date' columns	value = 0 if the cab booking was done within 24 hrs; value = 1 if the cab booking was done before 1 day upto a month
online_booking	categorical variable;	Value = 1 if booking was done on desktop website
mobile_site_booking	categorical variable;	Value = 1 if booking was done on mobile website

Exhibit 2: Sample Input for the data model

online_booking	mobile_site_booking	Is Customer Cancell	Booking Day_Of_Week	Pickup_Day_Of_Week	Binned Pick Up Month	Binned Booking Month	Binned_Booking_Time_Of_Day	Binned_Pickup_Time_Of_Day	Is_Booking_Within_A_Day
0	0	1	7	7	8	8	4	4	0
0	0	0	2	2	6	6	4	4	0
0	0	0	3	3	7	7	4	4	0
0	0	0	7	7	11	11	1	1	0
0	0	0	6	6	2	2	5	5	0

Exhibit 3: Classification Tree



Validation Data scoring - Summary Report (Using Full Tree)

Cut off Prob.Val. for Success (Updatable)	0.2

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	0	2072
0	0	17904

Error Report			
Class	# Cases	# Errors	% Error
1	2072	2072	100.00
0	17904	0	0.00
Overall	19976	2072	10.37

Exhibit 4: Output from KNN method

Test Data scoring - Summary Report (Using Full Tree)

Cut off Prob.Val. for Success (Updatable)	0.2

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	0	1369
0	0	11953

Error Report			
Class	# Cases	# Errors	% Error
1	1369	1369	100.00
0	11953	0	0.00
Overall	13322	1369	10.28

Validation Data scoring - Summary Report (for k=1)

Cut off Prob.Val. for Success (Updatable)	0.2

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	497	1575
0	2693	15211

Error Report			
Class	# Cases	# Errors	% Error
1	2072	1575	76.01
0	17904	2693	15.04
Overall	19976	4268	21.37

Test Data scoring - Summary Report (for k=1)

Cut off Prob.Val. for Success (Updatable)	0.2

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	347	1022
0	1793	10160

Error Report			
Class	# Cases	# Errors	% Error
1	1369	1022	74.65
0	11953	1793	15.00
Overall	13322	2815	21.13

Exhibit 5: Output from Naïve Bayes

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.2

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	348	1724
0	1232	16672

Error Report			
Class	# Cases	# Errors	% Error
1	2072	1724	83.20
0	17904	1232	6.88
Overall	19976	2956	14.80

Test Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.2

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	252	1117
0	873	11080

Error Report			
Class	# Cases	# Errors	% Error
1	1369	1117	81.59
0	11953	873	7.30
Overall	13322	1990	14.94

Exhibit 6: Ensemble method

Column Snapshot:

Predicted Class	Actual Class	Prob. for 1 (succes	Predicted Class	Actual Clas	Prob. for 1 (succes	Predicted Class	Actual Class	Prob. for 1 (succes	Ensemble 1	Ensemble 1 Predicti
1	0	0.0769702	0	0	0	1	0	0.0766283	0.051199	0
1	0	0.0790662	0	0	0.028571	1	0	0.0851761	0.064271	0

Count of Actual Class	Column Labels		
Row Labels	0	1	Grand Total
0	9836	2117	11953
1	895	474	1369
Grand Total	**10731**	**2591**	**13322**
Error Report			
Class	# Cases	# Errors	% Error
1	11953	2117	17.71
0	1369	895	65.38
Overall	13322	3012	22.61