Group A-4

Submitted by:

Arpit Gupta: 61410435
Mandeep Sandhu: 61410109
Manoo Kapoor: 61410817
Rishiraj Shrawat: 61410084
Udit Lekhi: 61410461

[SHARE AND REACH ANYWHERE]

Use of analytics to help provide shared cab service alternative to commuters plying on predicted congest routes

Contents

Executive Summary

'Cab sharing' is a well known concept of public transport service that can enable people to use taxi services at low cost. Primarily this service is targeted to achieve two broad goals. One is availability of an economically viable cab service option to areas which are poorly serviced by public transport wherein the only other alternative would be a high cost cab service. Shared cabs could cater to those commuters who travel frequently on highly congested routes and often face difficulties because of unavailability of good transport facility. Secondly cab sharing facility can help the cab service provider maximize revenue by catering any untapped customers, reducing per trip cost and avoiding any customers due to unavailability of cabs.

With roaring fuel prices and the need to maintain a high fleet of cabs several financial and operational challenges are faced by cab service providers including our client "Your Cabs". The biggest challenge a cab service provider faces is to maintain the right tradeoff between capturing maximum consumer surplus from the customers to whom cab service can be provided and between losing out on revenues from customers either due to unavailability of cabs or due to offered rate cards which are higher than customer's willingness to pay.

Also another operational issue which cab service providers face is variability of demand. Different routes at different time slots of the day have varied demand for the cabs. A model which can predict approximate number of cabs required from a particular route at a particular time would help cab service providers distribute or allocate resources in a more informed way.

Exhaustive data set that was made available provided us with data pertaining to details like the number of cabs booked from a particular area to a particular destination. Analysis of the data that was made available helped us gather insights such as routes such as (Airport, Whitefield, Marathahalli) that witness high traffic. Further analysis of sub set of the above data provides details showing a definite pattern in how bookings are done on a particular day or a particular time slot within a day.

Visualization of the data followed by application of data analytics such as classification techniques 'KNN' and 'Naïve-Bayes' would help us predict the demand levels among different routes. Further this prediction if compared to a *pre decided demand limit* can help cab service providers locate all congested routes that have a possibility of optimizing per trip cost by offering cab sharing.

With the designed data model, help desk personnel will have necessary data points to offer customers requiring a cab service the option to avail shared cab offered at lesser price (if the data model predicts high demand on the customer specified route/day and time). Also the prediction model can help cab service reduce operational costs by exploiting the possibility of offering shared cabs service and by maintaining an optimum cab fleet for a particular route thereby reducing the overall maintenance cost.

Detailed Report

Problem description:

Business Goal:

Our prime business goal is to help increase revenue for our client. With the usage of the designed model our client would be able to increase revenue by

1. Ability to tap untapped customer base without incurring additional capital expenditure
2. Lower maintenance and operational cost per rupee revenue
3. Better resource(cab) allocation

We intend to achieve this by equipping our client with a predictive model that depending on the input parameters such as source, destination, day of travel , time of travel etc can predict highly whether a route falls under congested route category or not. The categorization is done depending on probability of number cab bookings made below or above a pre defined demand limit.

Ability to tap unmet needs: There is high fluctuation in cab demand for certain routes on a definite time which poses operational challenges and turning down of cab. Ability to predict these types of routes at the time of booking can equip our client to offer low priced shared cab service to customers. Although low priced, combined revenue per cab would be higher making it profitable and serve additional number of customer without additional capital expenditure.

Lower maintenance and operational cost per rupee revenue: Shared cab service offerings would result in lower maintenance cost and would also result in lower fuel cost per rupee revenue from the customers. This makes our client service more attractive to cab drivers.

Resource allocation: Upfront prediction would also allow more optimized resource allocation

Analytics/Data Mining Goal:

We would like to adopt the approach wherein we conduct a supervised learning on the available data set. The output column of interest provides a prediction whether a particular combination of route, day and time of travel falls under the category of congested or uncongested route. The prediction would be determined based on a predetermined cutoff demand limit.

Data description:

For our analysis, we are primarily concerned with details of the record that pertain to the cab routes and time of the travel. We also decided to investigate any potential effects of mode of booking (online, mobile app, call center) and travel type. We do not require details regarding booking ID.
After preparing the relevant data (removing records with missing entries) and clustering records based on source-destination routes, the size of the data was reduced to 10,818 rows and 15 columns.
The specific attributes relevant to the analysis are as below:

Input Variables

- Cab route attributes:
  - Route Source Clusters: From_ID_1, From_ID_2, From_ID_3, From_ID_4, From_ID_5
  - Route Destination Clusters: To_ID_1, To_ID_2, To_ID_3, To_ID_4, To_ID_5
- Travel time attributes:
  - Day of the week

  - Binned Hour
- Misc. Details
  - Booking type
  - travel_type_id

Output Variables:

- Decision to provide shared service or not
  - Shared or Not?

Refer to appendix for sample records

The attributes described above are all derived from the raw data provided. The preparation of the same is described below.

Data Preparation:

We performed the following actions on the initial raw data

1. **Data Cleaning:** Since our analysis is primarily based on route congestion identification, we removed all entries which had NULL/missing entries in '*from_lat*', '*from_long*', '*to_lat*' and '*to_long*'.
2. **Extraction of travel time details:** From the '*from_booking'* attribute, we extracted the Week day and hour of the day to create 2 new attributes: '*Day* of the week' and '*Booking_Hour* '. Booking_Hour was binned into 3-hour slots as '*Binned_Hour_1/2/3/4/5/6/7/8'*.
3. **Clustering of locations:** Using the latitude and longitude details in '*from_lat/long'* and '*to_lat/long*'', we clustered all the locations based on their geographical locations in the city. Hence, we formed 5 clusters with locations in each cluster being geographically close by. The *from/to* details in the records were represented with a dummy variable: *From_ID_1/2/3/4/5* and *To_ID_1/2/3/4/5*.
4. **Clustering of records:** All records with the same values in all attributes were collapsed into a single record. A new record *'Total bookings'* was added which was equal to the number of records that were collapsed into a single record.

Output variable added: Attribute '*Shared or Not?*' was added. If the value in '*Total bookings'* was above a pre-decided value (15), the value was set to 1, else 0. This means that if the number of bookings on a particular route in a 3 hour slot was beyond 15, there is a potential to provide a shared cab service.

Data Mining Solution

Since the outcome variable (Shared or Not?) was only 10% of the total dataset we created, we had to oversample the data to be able to run the model for predicting the output variable.

Step -1: **Partitioning with Oversampling**: We partitioned with oversampling ensuring 30% success of allowing sharing a cab in the training data and allowed test data to have 30% of the validation data set.
Step -2: **KNN Algorithm** : We ran the K-Nearest algorithm to identify to best k possible and noticed that k=19 came out to be the best k for which we ran the algorithm to classify and develop the predictive model for our perusal. We kept the success probability cutoff at 0.35. Basis the model our error rates came out to be as below:

Data Set	Error Rate(%)
Training	**33.54**
Validation	**39.45**
Test	**40.32**

Step -3: NNB Algorithm : We ran the NNB algorithm to build a predictive model on the output variable "Shared or not?" to identify if the model predicted any better than the KNN based model. We kept the success probability cutoff at 0.35 here as well. Post the analysis we noted the following error rates:

Data Set	Error Rate(%)
Training	**33.82**
Validation	**35.09**
Test	**36.81**

Step-4: Ensemble :Since from both the models we saw the error rate to come out as quite high, we considered combining the two models to see if the error rate improves on the data set. For doing the same, we took out the test data set for consideration and used the average probability as an indicator for the ensemble. We observed that the accuracy for the test data improved quite significantly as it become **26.5%** enabling us to look at a better predictive model.

Performance Evaluation

The performance of the model would be decided on the extent of additional revenue that can be generated by our client. We understand that there would be errors in prediction and there would be two types of errors

1. Misclassification of congested areas as non congested areas (Type one error)
2. Misclassification of non congested areas as congested areas (Type two error)

As per the confusion matrix, assumptions taken & the profit calculation shown (refer appendix 3) we can see that even when we account for the revenue loss due to type 2 error our client would be able to break even. In addition to this, there will be approximate capacity of 40 cabs that would still be available to cater to extra capacity thereby enhancing revenue. Revenue loss would be directly proportional to no. of type 2 errors while revenue gain would be dependent on correct classification of congested route.

Conclusion

Advantages: We have used Naïve Bayes which is simple to model and quicker than descriptive models like logistic regression and thus requires lesser training data. In addition to this the model was also formulated using KNN classification algorithm which has its own advantages such as the model can be updated at a very little cost. In addition to this we have used the concept of ensemble wherein we have merged two models to getter better predictive prediction.

Limitation: Since we are using KNN algorithm, prediction of each new instance involves comparison with every other record in the database. Although modern computing systems are robust, there is sensitivity to the increase in number of records in the database. Another limitation is associated with the usage of Naive Bayes algorithm. The prediction for a new record which has no matching record in the model assumes a zero probability for its outcome. Hence, if there is some event happening in the city at a place where there is no traffic otherwise, the algorithm will predict zero forthcoming demands.

Also the clustered longitudes and latitudes are from exiting data sets and new areas have to be added on going basis

Operational Recommendations: When a booking comes, the operator needs to classify the "From" and "To" latitudes and longitudes on the basis of the geographical clusters developed in the model. Post this, the KNN algorithm and NNB algorithm would be run to predict if shared service can be provided(based on level of demand). This should be implemented at mobile, calling and web applications for it to be successful. The price for sharing a cab should be modulated to include the expected breakeven

Appendix

Sample Records:

Row Id.	Selected variables															
	From_ID_1	From_ID_2	From_ID_3	From_ID_4	From_ID_5	To_ID_1	To_ID_2	To_ID_3	To_ID_4	To_ID_5	Day of the week	Binned_Hour	Booking type	travel_type_id	vehicle_model_id	Shared or Not?
7219	0	0	0	1	0	0	0	1	0	0	7	8	2	2	64	0
3642	0	0	1	0	0	1	0	0	0	0	5	2	1	2	12	0
8833	0	0	0	0	1	0	1	0	0	0	3	1	3	2	28	0
8031	0	0	0	1	0	0	0	0	0	1	4	3	1	2	24	0
4028	0	0	1	0	0	0	1	0	0	0	3	8	1	2	85	0

Partition

Partitioning Method	Partitioning with oversampling
Random Seed	12345
# training rows	1783
# validation rows	3787
# test rows	1622
Selected output variable	Shared or Not?
% Success in Training data	30
% Validation data taken away as test	30
% Success in original data set	9.890922537

Row Id.	Selected variables															
	From_ID_1	From_ID_2	From_ID_3	From_ID_4	From_ID_5	To_ID_1	To_ID_2	To_ID_3	To_ID_4	To_ID_5	Day of the week	Binned_Hour	Booking type	travel_type_id	vehicle_model_id	Shared or Not?

KNN

Best K

Value of k	% Error Training	% Error Validation	
1	12.39	37.07	
2	20.58	40.69	
3	20.19	29.44	
4	23.89	32.19	
5	23.16	26.17	
6	24.79	29.10	
7	24.40	24.00	
8	25.29	26.22	
9	25.57	22.02	
10	26.58	23.82	
11	27.20	20.41	
12	27.37	21.73	
13	27.48	18.99	
14	27.99	20.10	
15	27.82	17.90	
16	28.27	19.04	
17	28.66	16.90	
18	28.88	17.53	
19	28.72	15.82	<--- Best k

Training Data

Training Data scoring - Summary Report (for k=19)

Cut off Prob.Val. for Success (Updatable)	0.35	(Updating the value here w ill NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	379	156
0	442	806

Error Report			
Class	# Cases	# Errors	% Error
1	535	156	29.16
0	1248	442	35.42
Overall	1783	598	33.54

Validation Data

Validation Data scoring - Summary Report (for k=19)

Cut off Prob.Val. for Success (Updatable)	0.35	(Updating the value here w ill NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	236	139
0	1355	2057

Error Report			
Class	# Cases	# Errors	% Error
1	375	139	37.07
0	3412	1355	39.71
Overall	3787	1494	39.45

Test Data

Test Data scoring - Summary Report (for k=19)

Cut off Prob.Val. for Success (Updatable)	0.35	(Updating the value here w ill NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	111	49
0	605	857

Error Report			
Class	# Cases	# Errors	% Error
1	160	49	30.63
0	1462	605	41.38
Overall	1622	654	40.32

NNB

Training Data

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.35	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	370	165
0	438	810

Error Report			
Class	# Cases	# Errors	% Error
1	535	165	30.84
0	1248	438	35.10
Overall	1783	603	33.82

Validation Data

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.35	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	250	125
0	1204	2208

Error Report			
Class	# Cases	# Errors	% Error
1	375	125	33.33
0	3412	1204	35.29
Overall	3787	1329	35.09

Test Data

Test Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.35	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	113	47
0	550	912

Error Report			
Class	# Cases	# Errors	% Error
1	160	47	29.38
0	1462	550	37.62
Overall	1622	597	36.81

Ensemble

Classification Confusion Matrix		
	Actual Class	
Predicted Class	0	1
0	1114	82
1	348	78

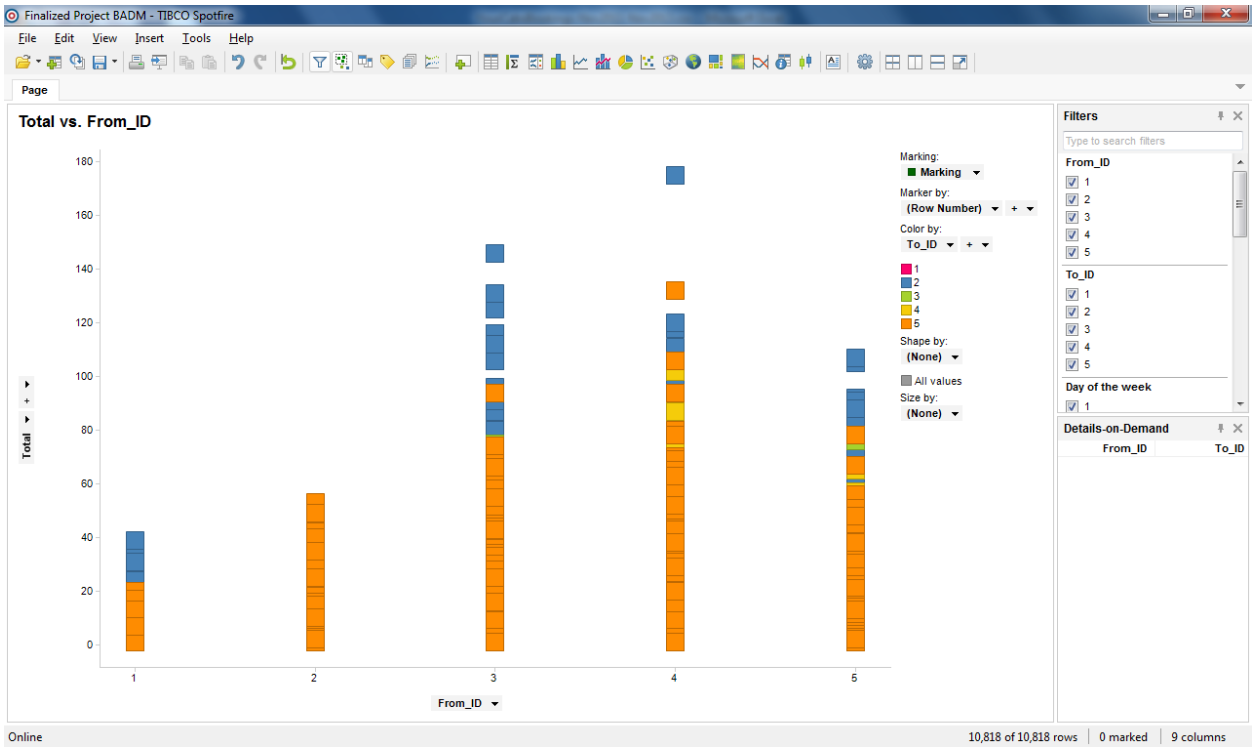
Performance Evaluation

Assumptions	
Revenue per trip	500
Cost incurred per trip	250
Profit Per trip	250
Total no of cab bookings	1622
Total profit	405500

Actual=0 Predicted=0		Actual=1 Predicted=0	
Revenue per trip	500	Revenue per trip	500
Cost incurred per trip	250	Cost incurred per trip	250
Profit Per trip	250	Profit per trip	250
Total no of cab bookings	1114	Total no of cab booking	82
Total profit	278500	Total profit	20500
Actual=0 Predicted=1		Actual=1 Predicted=1	
Revenue per trip	400	Revenue per trip	800
Cost incurred per trip	250	Cost incurred per trip	250
Profit Per trip	150	Profit per trip	550
Total no of cab bookings	348	Half the no of cabs	39
Total profit	87000	Total profit	21450
Grand total	407450		

Data Visualisations:

The below chart depicts number of bookings from each of the 5 source clusters color sorted as per destination clusters. The high density of orange cluster(cluster 5) suggests that a lot of bookings cater to cluster 5.



The below pie chart describes the distribution of bookings to each destination cluster as per the binned hour of the day. The pie chart suggests near equitable distribution to all clusters on all days. Hence, we cannot conclude a concentration of cabs to any cluster at a particular hour of the day.