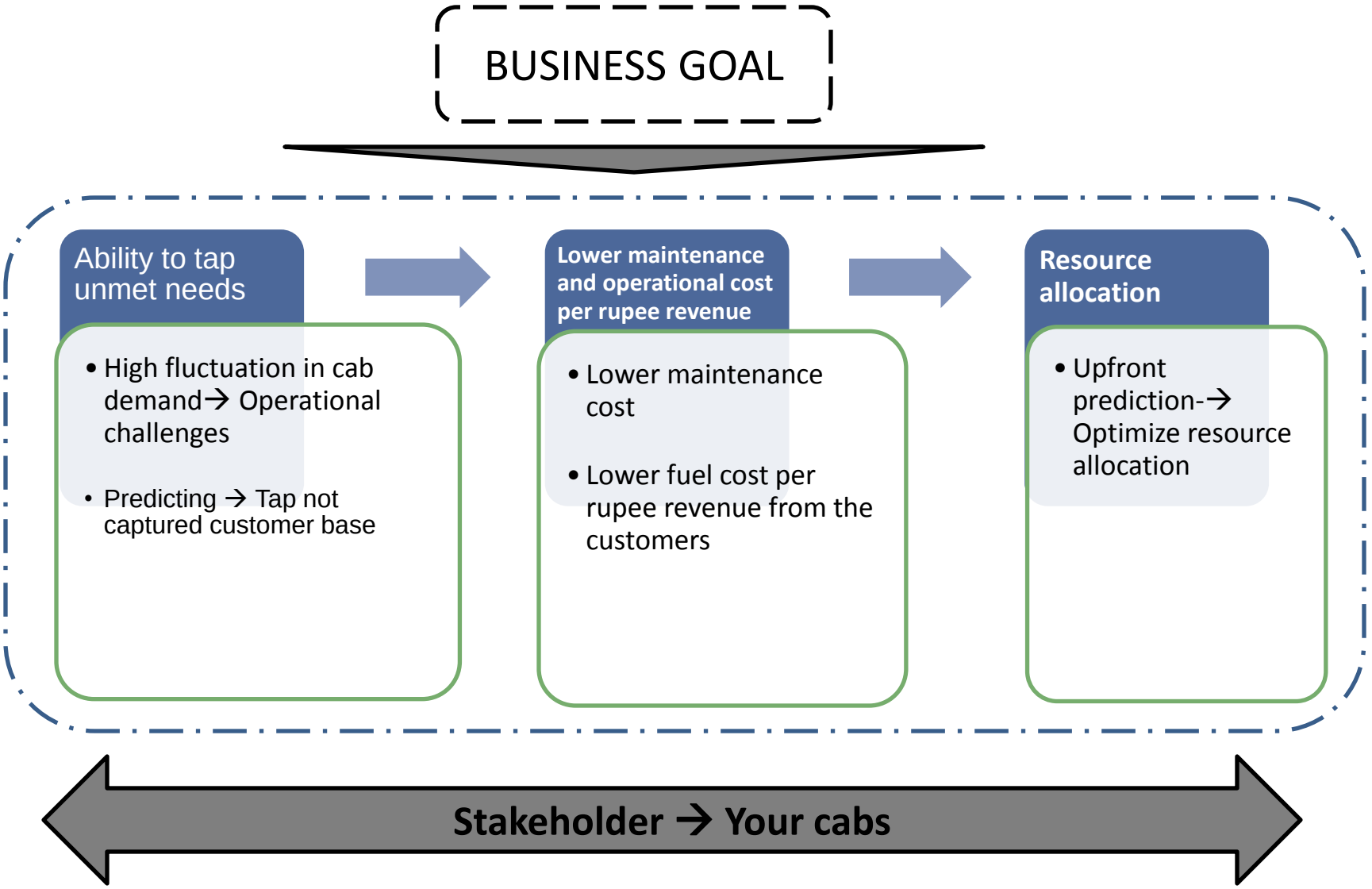
Share and Reach anywhere

Group A 4

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

Business Problem



Data-Mining Problem

PREDICTIVE
The output column predicts whether a particular combination of route & time falls under congested route

DATA MINING GOAL

SUPERVISED
Conducted a supervised learning on the available dataset

CUSTOMER TOUCH POINTS

A customer calls to book a cab

System takes the route & timing info as an input

Depending on the provided information check for congested /non congested route

Based on the output, the user is offered a shared cab option with differential charges

Data description

Input Variable

- Cab route attributes
  - Route Source Clusters
  - Route Destination Clusters
- Travel time attributes
  - Day of the week
  - Binned Hour
- Misc. Details
  - Booking type
  - travel_type_id

Output Variable

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

Data Preparation

1. Data Cleaning
2. Extraction of travel time details
3. Clustering of locations
4. Clustering of records
5. Output variable added

Sample Row details

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

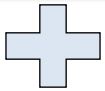
Methods

Partitioning with Oversampling

-Post Data Preparation, only few of data records
-Output Variable (“Shared or Not?”) occurrence only 10% approx
-Oversampled with 30% of success in training data

K-Nearest

-Best Possible K Identified as 19
-Output Variable as “Shared or Not?” with 1 indicating to be shared and 0 as not to be shared
-Chose success probability as 0.35
-Error Rates – Test (40.32%), Training(33.54%), Validation(39.45%)



Naïve Bayes

-Chose success probability as 0.35
-Output Variable as “Shared or Not?” with 1 indicating to be shared and 0 as not to be shared
-Error Rates – Test (36.81%), Training(33.82%), Validation(35.09%)

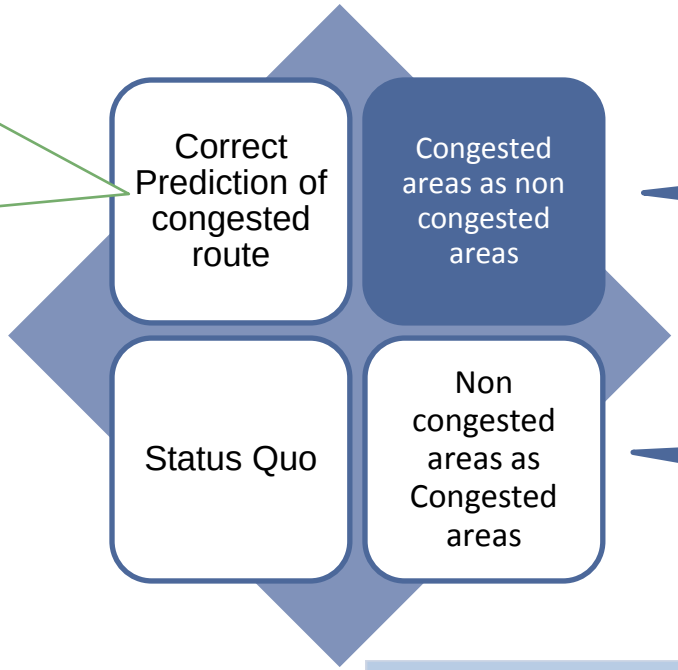
Ensemble

-Combined KNN and NNB Model
-Used the average probability as an indicator for finalizing the ensemble prediction
-Error Rate for test data reduced drastically to (26.5%)

Data Created for the purpose of the analytic goal

Evaluation

Higher correctness of this prediction will lead to higher revenues



Costly error resulting offer of shared cab to single customer

Loss of Opportunity cost

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

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		

Recommendations