The Impact of Data Mining for Insights in Travel and Tourism Industry

Prabal Mahanta and Suhas Sudheendra

***Abstract*—The main interest for a tourism industry lies in analyzing the financial side like revenue per room and traveler expenses. The traveler analysis is provided in most of the analytics but if we can view it from the trip perspective then it will provide more meaningful insight. How a traveler is affected by economy, policies, media and other factors are few of the insightful parameters that we can consider. The concept deals in the area of tourism analytics from the big data perspective and how handling data from the tourism perspective can be made easy. The concept refers to the earnings and the households of the destination and churns the mathematical model to come up with a score for each trip a traveler makes. Based on the model, we also propose architecture for handling the data and provide analytics on the fly. The main aim is to study the tourism industry and come up with various insights to the data and the issues related to the same.**

The gamified aspect of the industry is also one of the key features to study which we take into consideration and apply forecasting and prediction techniques to see the impact on the revenue or contribution in the future.

***Index Terms*—In-memory computing, tourism, analytics, gamification.**

I. Introduction

Roughly 3% of world's GDP is through tourism, which is substantial [1]. Like all other industries this can be improved. There are organizations which currently analyze the tourism data based on the number of tourists arriving, whether they are inbound or outbound and their total expenditure on various infrastructures. Due to recent economic conditions spending in general is reduced but it has not affected much on tourists spending. By 2023, contribution from tourism will be around US$ 10.5 trillion in GDP. This leaves us with a tremendous task of uplifting the infrastructure and changes in tourism policies. But to understand the impact of each we need analyze the data and come up with right strategies so that we can meet or exceed the forecasted amount.

Tourism brings in opportunities for employment and in this regard it tries to compensate for the unemployment and poverty in the region. It is very important to understand the pattern of tourism in a region to best address in reforming the policies to bring more tourist inflow.

The data that comes from the field of tourism is distributed and dispersed. To analyze such a data it is important to have an architecture and tool to visualize the insights. Here is where we come up with the concepts of utilizing the in-memory technology along with various data mining techniques.

One of the key factors here is to understand the spending pattern. The region of interest always tends to invest on infrastructure, transport, other various facilities to attract tourist to that place. The result of which are changes in policies in the region to procure more funds to accomplish and maintain the same. To make this process successful various surveys and information gathering is done. The part where this may lead to wrong analysis is the nature of this data being random in nature.

The aim is to propose a model for the industry where the specific regional heads will have proper insights to the data without having to reiterate the processes and to conclude on policies without substantial confidence.

With the in-memory analytics [2] concept we get faster and accurate forecasts and insights which will help the decision makers to better plan for the industry and also the region of interest leading to lot of opportunities in the sectors of employment and hence contributing to the GDP.

Since we have a revenue generating case study, we can say that traditionally it often happens that Revenue models often assume that as the date of activity nears the consumers are more tending to pay for the same (Raeside, 1997) [3].

In this industry often the customer's decisions are influenced by the likelihood of getting a good deal than the present offerings and often it is also greatly influenced by sellout of the deal, this is a common observation during the holiday season in the tourism industry.

The deals and other offers that attract the traveler is also the result of recent advances in the concept of gamification in travel [4]. The aim of our on-going research is to drive the concept of gamification to the travel industry from the tourism department point of view which will help them compete better among one another [5], [6].

This paper tries to come up with the concept of Propensity score matching (PSM) [7] and can be used across all the observational studies to reduce the potential selection bias due to random data sets. The process technically comprises of the score estimation (propensity), evaluation based on the matching process [8], [9].

II. Present Research

The field of study is tourism and the data in this field is vast and the aggregated data is exposed to the public for inference which is not a good measure for detailed insights. The measure of the data bias is very difficult but the same can be categorized as risk and since it is will contribute to the majority then it will be easier to adjust for minority of events

Manuscript received June 9, 2013; revised September 16, 2013.

Prabal Mahanta and Suhas Sudheendra are with SAP Labs Pvt Ltd Bangalore, India (e-mail: p.mahanta@sap.com, suhas.s@sap.com)

which can be derived for per individual and here individual can be considered as a variable and since there is a disparate distinction of parameters of money and time, propensity scores will help understand the problems and insights better for observed covariates.

The bias is not always necessary to negotiate in terms of risk and the errors in smaller samples cannot be avoided so the concept consider this as a constant for each variable and after the training of the observation samples the constant is varied for the iterations [10].

Since the data here is diverse and from different sources the data from survey, internet and other mode of data collection may not converge and this leads to perception of customers leading to negative bias towards the region in discussion [11], [12].

The issue is also due to random data in the sets which is required to be converging to a bias variable so that the sampling error reduces and random variation is also reduced. All the surveys are affected by the coverage error which is the issue with the survey design and the target consumer is not kept in the design reach. E.g. few people will rate a rough terrain hill station after they return from the place and the terrain prevents any internet activities and thus one who had unwanted incidents can lead to bias survey results [13].

The gamification aspect of the industry of travel is that the industry revolves around conditions and this conditional probability for each traveler will help us to form groups which will be analytically important to understand the categories of travel and we also can apply the concept to adjust the regression model using a covariate (Dimopoulos *et al*., 2008) [14].

So we consider two countries and 8 most likely influencers. We have C1 and C2 and each have parameters P1 to P3.

The architecture of our model can be seen in Fig.1.

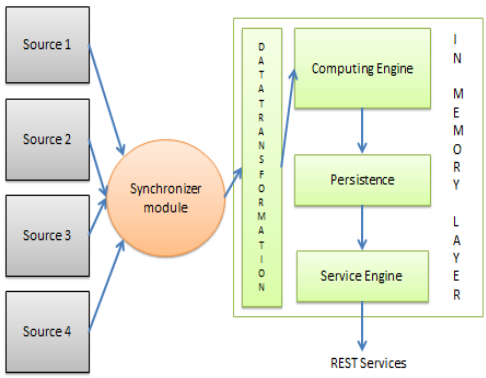

Fig. 1. Architecture of case study.

We consider various sources which will have data in various formats and structure. The Synchronizer module and the data transformation module works closely where we crawl and restructure the data in unified format and also cleanse the data for any redundant data which may be there in the country specific survey and open source data.

The in-memory layer comprises of the data transformation layer which works in sync with the synchronizer module, Computing Engine which applies the models to the data and the Persistence module helps in persisting the results in storage and the also feeds the same to the Service Engine which produces the REST services for the users to consume.

Going back to the experiment in study, we have countries C1 and C2 along with 8 deciding parameters.

Now we collect the open source data including survey, annual results and then observe the trend that influences the employment.

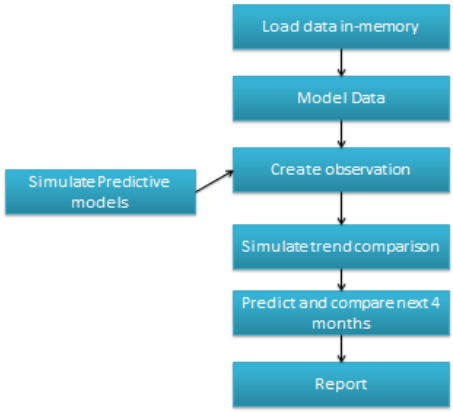

Fig. 2. Algorithm for processing data.

The data (see Fig. 2) we considered had to be modelled for the parameters and based on the same a propensity score is given to each which we can adjust according to the rules like:

1) Parameter can be ignored given the value of the score (Propensity score).
2) Cleansed data which we already perform during initial phase of data sync.
3) Observed covariates represent unobserved covariates.
4) Parameter tweaks may not affect the covariates.

The concept is very dominant in health care industry where impact of drugs on patient groups is measured.

Now we observe data for the parameters P1, P2 and P3 based on the modeling on the survey data for each countries C1 and C2 (See Fig. 3 and Fig. 4).

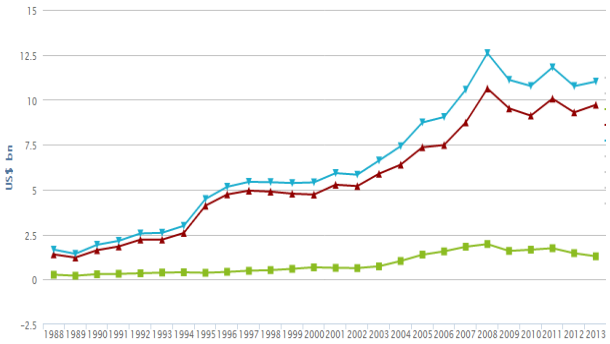

Fig. 3. Country 1 Analysis for yearly P1, P2 and P3 impact on the revenue.

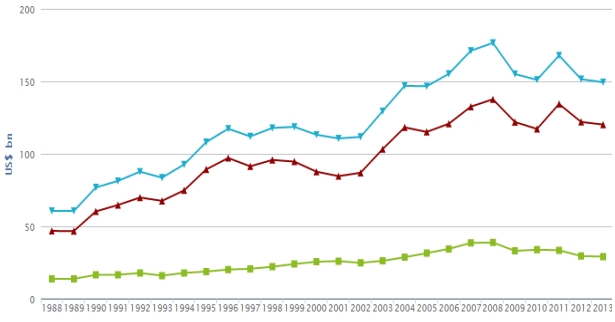

Fig. 4. Country 2 Analysis for yearly P1, P2 and P3 impact on the revenue.

We observe that both are having huge gaps in terms of revenues as parameters P2 and P3 impact 10 times for the country C2 than the country C1.

We then investigated by adjusting the propensity scores for parameter P2 for both countries C1 and C2 for 2 different aspects.

This experiment was led by gamified data for each country for parameter P2. Now if we observe P2 there were two trends that came out of the data which were like indirect and direct contribution to the revenue which can be seen in the Fig. 5 and Fig. 6.

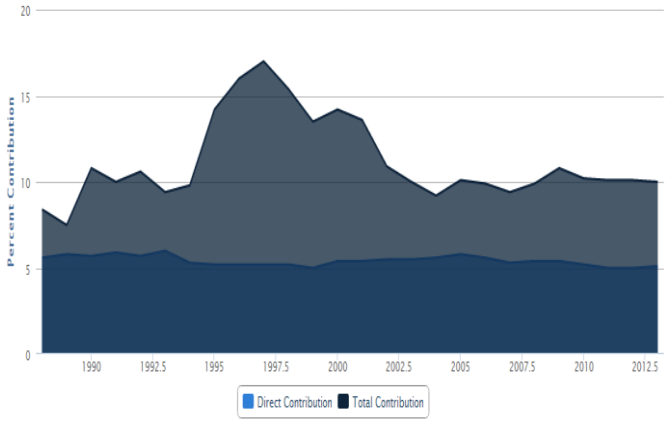

Fig. 5. Country 1 Analysis for yearly P2 impact on the revenue for direct and indirect mode.

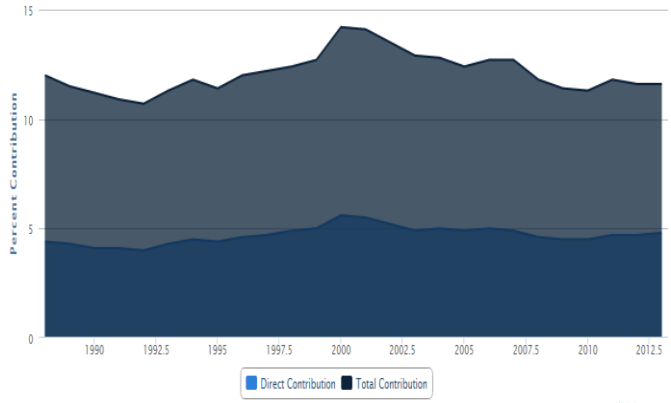

Fig. 6. Country 2 Analysis for yearly P2 impact on the revenue for direct and indirect mode.

Now we consider the gamified data which means the impact of the gamification during the stay of the tourists and extrapolated the impacts for parameter P2 for next 12 years by using predictive analytics.

We observed that P2's indirect and direct contribution can merge after certain period and this may lead to enhanced tourist flow in the country.

This came across as a very interesting observation from the industry perspective and this can help industries similar to this one which has conditional parameters to improvise their business through gamification and preventive analytics using prediction.

The results for the gamified analysis is shown in Fig. 7 and Fig. 8.

There are various pattern of survey observed while applying the concepts of the gamified simulation on the data. The experiments also done to simulate random regional specific population with parameters which decide the propensity to travel based on training sets prepared for the inflow and outflow of population based on the historical data.

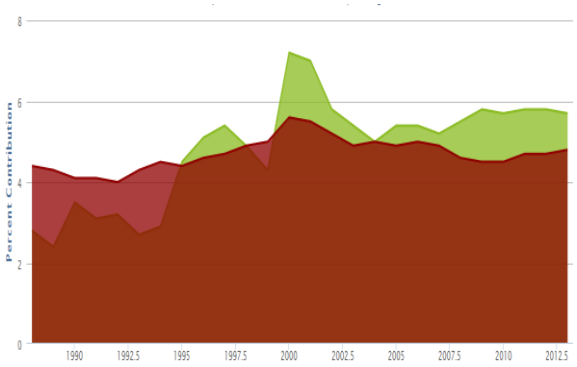

Fig. 7. Country1 Analysis for P2 impact (Gamified).

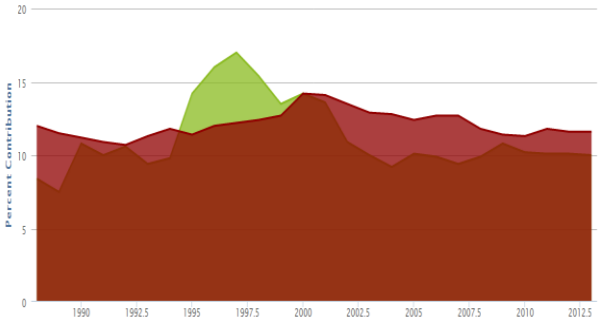

Fig. 8. Country2 Analysis for P2 impact (Gamified).

The experiment depicts the population flow pattern from the region of origin to the region of interest based on the parameter values which also was varied with a chaos factor and this chaos factor increased and decreased the propensity factor which signified theoretically a gamified factor which affected the flow of pattern of the tourist from region of origin to region of interest in varied time varied instances which signified the season and this parameter shows that with aggressive manipulation of the interest factors the tourism sites can draw tourists all year long.

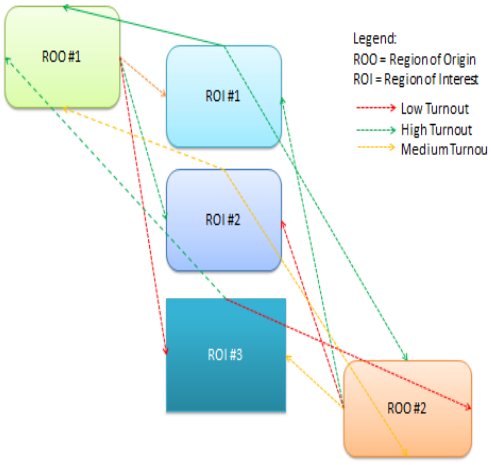

Fig. 9. Experiment for a particular season for outflow and inflow simulation of tourists from region of origin to region of interests.

As shown in Fig. 9 and Fig. 10, the experiment was able to convert the low turnout regions to medium turnout region of interests and the stats can be seen in Fig. 11.

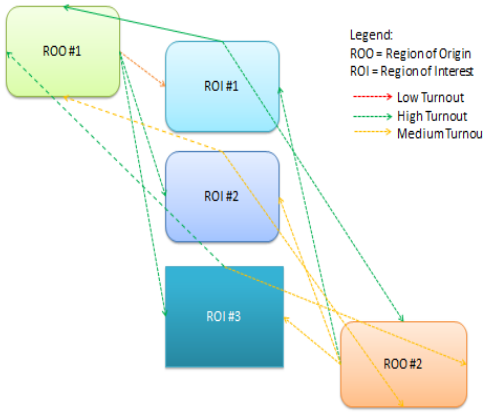

Fig. 10. Experiment for a particular season for outflow and inflow simulation of tourists from region of origin to region of interests with Gamified factor resulting in increase of turnout from Low to Medium.

GAME FACTOR = 0	FLOW OF TOURIST PER 100		
	ROI 1	ROI 2	ROI 3
ROO 1	22	68	10
ROO 2	56	12	32
GAME FACTOR = 1	FLOW OF TOURIST PER 100		
	ROI 1	ROI 2	ROI 3
ROO 1	22	68	10
ROO 2	56	12	32

Fig. 11. Stats for the flow of tourists in simulated environment.

III. Conclusion

The data plays a vital role in any industry and the people opinion is a great influencer for the revenue generation policies of specific region of interest. There are often biases in these data which we have to negotiate to come up with different and deeper insights to any industry. This research is still in development stage and we are trying to see how we can improvise on the machine learning to get the simulation of real-time analysis for the tourist flow and the impact of their activities on the revenue of the region of interest.

IV. Application in Other Industries

The similar concepts can be applied to healthcare and automotive industries and in present times the healthcare industry already has research going on these terms. We can also map this concept to the retail industry and this will help understand the gamification factor in the industry and then the perspectives can help in greater insights.

Acknowledgment

The authors would like to express gratitude for all the support and motivation provided by Mr. Ganapathy Subramanian, Vice President, TIP DNA (CE&SP), SAP Labs India, Bangalore. We acknowledge his support for encouragement and guidance during this research.

References

[1] J. Falconi, "Measuring the economic contributions of tourism: A proposal for some basic indicators," *Enzo Paci Papers on Measuring the Economic Significance of Tourism*, vol. 3, 2003.
[2] SAP HANA. (July 2013). [Online]. Available: http://www.sap.com/hana/hana-database/.
[3] R. Raeside, "In yield management: Strategies for the service industries," in *Quantitative Methods*, I. Yeoman. and A. Ingold, Ed. London: Cassell, 1997.
[4] T. Bekker, J. Sturm, and E. Barakova, "Designing for social interaction through physical play," *Personal and Ubiquitous Computing*, vol. 14, no. 5, pp. 281-283, 2010.
[5] C. Crumlish and E. Malone, *Designing Social Interfaces: Principles, Patterns, and Practices for Improving the User Experience*, 1st ed. Sebastopol, 2009.
[6] (2013,14 July). Gamification in Travel and Transport. [Online]. Available: https://www.travelandtransport.com/loyalty/programs/gamification /
[7] D. Rubin and N. Thomas, "Characterizing the effect of matching using linear propensity score methods with normal distributions," *Biometrika*, vol. 79, no. 4, pp. 797-809, 1992.
[8] A. Abadie and G. W. Imbens, "Large sample properties of matching estimators for average treatment effects," *Econometrica* , vol. 74, no. 1, pp. 235-267, 2006.
[9] A. S. Goldberger, *A Course in Econometrics*, Cambridge: Harvard University Press, 1991.
[10] J. Armoogum, M. Herry, J. L. Madre, and J. Polak, "Sampling and weighting schemes for travel diaries: Review of issues and possibilities," *Methods for European Surveys of Travel Behaviour*, no. 6, 1996.
[11] J. Barton, "Multi-Household procedures for social surveys," *Survey Methodology Bulletin*, vol. 40, 1997.
[12] J. C. Deville and C. E. Sarndal, "Calibration estimators in survey sampling," *Journal of the American Statistical Association*, vol. 87, pp. 376-382, 1992.
[13] D. Elliot and S. Bruce, "Person vs household weighting on the UK Labour Force Survey," *Survey Methodology Bulletin*, vol. 43, pp.43-52, 1998.
[14] K. Dimopoulos, G. P. Diller, E. Koltsida, A. Pijuan-Domenech, S. A. Papadopoulou, S. V. Babu-Narayan, T. V. Salukhe, M. F. Piepoli, P. A. Poole-Wilson, N. Best, D. P. Francis, and M. A. Gatzoulis, "Prevalence, predictors, and prognostic value of renal dysfunction in adults with congenital heart disease," *Circulation*, vol. 117, no. 18, pp. 2320-2328, 2008

Prabal Mahanta was born in the city of Guwahati, Assam, India in 1985.

Mr. Mahanta holds a Master's degree in Information Technology (M.Tech) with majors in Embedded Systems from International Institute of Information Technology (IIIT), Bangalore, Karnataka, India. He is currently working as a Developer at SAP Labs, Bangalore, India. He has worked extensively on critical customer projects as well as research projects. He has a strong inclination for research and teaching. He has presented many papers in international and national conferences and his research interest lies in in-memory computing, cloud-infrastructure, chaos theory and real time simulations.

Suhas Sudheendra was born in the city of Mysore, Karnataka, India in the year 1983.

Mr. Suhas holds a Bachelor's degree in Computer Science and Engineering from Visvesvaraya Technological University Belgaum, Karnataka, India. He is currently working as a Senior Developer at SAP Labs, Bangalore, India. He has worked extensively on research projects from the beginning of his career at Infosys which includes performance engineering and application optimization. Current areas of interest include in memory computing and real time analytics.