Volume 4 Issue 1

ISSN: 2321-1857

International Journal of

Business Analytics & Intelligence

April 2016

A Publication of Publishing India Group

International Journal of Business Analytics & Intelligence

Editorial Message

In the fourth year of its inception, IJBAI comes out in a new avatar with a clear vision and enriched contents to contribute to the future of analytics. Firstly, IJBAI will play a prominent role in sharing the application of analytics in industry. In the current issue, Ms. Madhumita Ghosh of IBM has contributed the business use case on how text analytics helps in generating brand perception insights. Secondly, we will also share in each of our issue a column or perspective on analytics from one of the renowned Professor of analytics. In the current issue, Professor Sudhir Voleti of ISB, Hyderabad has put forward his perspectives on the evolution of analytics. Thirdly, to publish the best of the research in academia, we have collaborated with multiple leading conferences across the world. In the current issue, we have published some of the selected papers from Business Analytics and Intelligence Conference held at IIM Bangalore. We are deeply indebted to Prof. Dinesh Kumar of IIM Bangalore on this regard.

We are extremely proud to have Dr. Beverly Wright, Ms. Madhumita Ghosh and Prof. Kevin Potcner in our editorial board. The new addition to the editorial board provides the much needed direction and support to make IJBAI the leading repository of knowledge in analytics. All of the above constructive steps are taken to delight our avid readers. Therefore, it would be great to have valuable feedback from our learned readers about the enriched version of IJBAI.

Best wishes,

Tuhin Chattopadhyay, Ph.D.
Editor-in-Chief
E-mail: dr.tuhin.chattopadhyay@gmail.com

Dated: 1st June, 2016

Place: Mumbai, India

International Journal of Business Analytics and Intelligence

Volume 4 Issue 1 April 2016

ISSN: 2321-1857

Case Study

'Text'-Mining Customer's View Point and Perceived Value About Brand
Madhumita Ghosh 1-4

Perspective

The Exponential Learning Curve of Analytics
Prof Sudhir Voleti 5-6

Research Papers

1. **Airline Revenue Management – Revenue Maximization Through Corporate Channel**
   Karthik V, Indranil Mitra 7-16
2. **Forecasting The Stock Market Values Using Hidden Markov Model**
   R. Sasikumar, A. Sheik Abdullah 17-21
3. **Claim Analytics across Multiple Insurance Lines of Business**
   Ravi Chandra Vemuri, Balaeswar Nookala, Ramakrishnan Chandrasekaran, Madhavi Kharkar, Sarita Rao 22-28
4. **Delay Prediction of Aircrafts Based on Health Monitoring Data**
   B. A. Dattaram, N. Madhusudanan 29-37
5. **Is Your New Product Really Boosting our Sales? An Econometric Model to Quantify the Cannibalization Effect**
   Vamse Goutam 38-44

Journal is available online at www.publishingindia.com

Article can be accessed online at http://www.publishingindia.com

'Text'-Mining Customer's View Point and Perceived Value About Brand

Madhumita Ghosh*

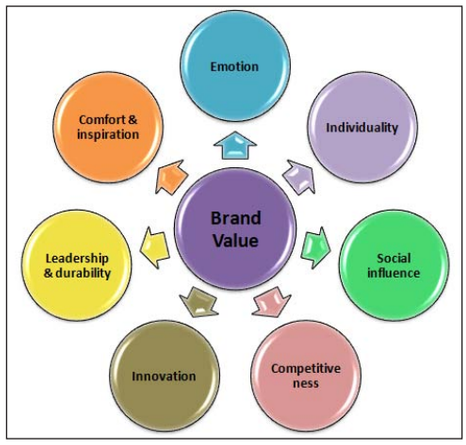
Abstract

This paper describes how text mining techniques can be applied in the analysis of consumer voice to gain useful and actionable business insights for marketers. The technique is illustrated via its application to understand Brand's perceived value of certain automobile brands. This case study shows the use of text mining techniques to understand brand's perception vis-a-vis competition from their opinion, sentiment and reactions. As the amount of online text increases, the demand for text classification to aid the analysis and management of text is increasing. Data acquisition in this case is not costly, information is rich in nature, classification of text can provide this information at low cost, but the classifiers themselves must be built with expensive human effort, or trained from texts which have themselves been manually classified. In this paper, we mention about a procedure of classifying text using the concept of association rule of data mining and correspondence analysis for Brand perception.

Voice of the customer analysis can have significant value for organizations looking to listen to and understand the customer's "voice" (e.g., from surveys, social media, complaints or web chat) to improve operations and help direct strategy. This approach can, ultimately, help improve customer satisfaction, Net Promoter Score (NPS) and loyalty while reducing churn and dormancy, thus increasing revenues. Consumers' Experience about a brand depends upon their expectations and engagement across touch-points of the brand. Assess Customer's Purchase, Usage & Service experience and mindset from social media as emerged touch point helps to understand Brand Imagery

Introduction

In today's cut throat competitive marketing, just because a brand wants to stand for greatness doesn't mean most of the people will convey great things about it. Brands are not just what they say about them, they are what consumers say they are. A brand's true identity lies in its 'perceived value'.



If marketers want to gauge what people think about their brands, there are a variety of (mostly academic) survey methodologies and feedback loops to utilize. Unfortunately, most brand research studies take too long to set-up and administer, and are not timely enough to optimize campaigns. These studies are very expensive and subject to significant survey biases, Social media being a growing touch point between consumer and brands, it is an effective platform to gauge their mindset by mining enormous texts. The honest impression & the most accurate composite of a brand's true identity seem to come from a consumer's first gut reaction to it and social sites are the better capture points, where consumers speak their mind.

The qualitative markers go beyond the typical gauges of brand awareness to encompass how consumers feel about a brand, how they think about it, react and talk about it, and interact with it. There are typically Eight trigger points or areas, which is essential to understand in perception studies.

* Practice Leader - Big Data & Advanced Analysis BA & Strategy - Global Business Services IBM, India.

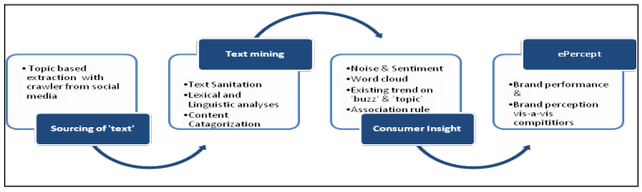
Techniques in text analytics balance existing data mining techniques to help make data richer, and insights more accurate. The technique is built on content management and text search, rather than databases and queries.

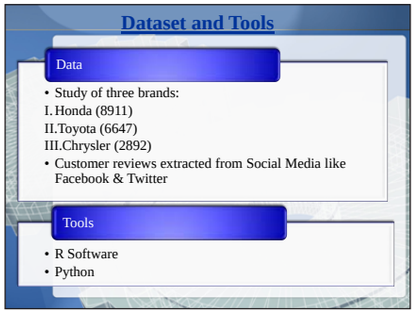
Approach

Keyword searches across various feed often result in the return of too much irrelevant information, which would take a long time to analyze. More sophisticated NLP approach was used, which enables to extract what customers are actually saying about a subject, as opposed to just mentioning a subject, Cluster together the different ways of saying the same thing as well as categorize populations (the text writers) on site according to their behaviours and opinions.

Three Automobile Brands were under study. Viz. Honda, Toyota & Chrysler. Customer reviews were extracted from social media sites with relevant sites and relevant key words.

The basic framework of the complete approach is depicted in this below picture:





Sourcing of Text: By deploying a propriety tool assess consumer comments from business pages on social media, other industry forums frequented by consumers, and websites pertaining to competitor's products. The information is collated and checked for any missing or incomplete parts. Connectors were deployed, tested, and fine-tuned to automate the data captured. Various 'key words' was tagged to refine the search and to receive more relevant textual information. It helps to minimise the junk at the basic level.

Below picture depicts certain example of texts fetched from social sites.

There were customer reviews for three brands namely Honda, Toyota and Chrysler given by the residents of North America. The corpus of these brands consisted of 8,911, 6,647 and 2,892 related reviews respectively, which was further formatted in database.

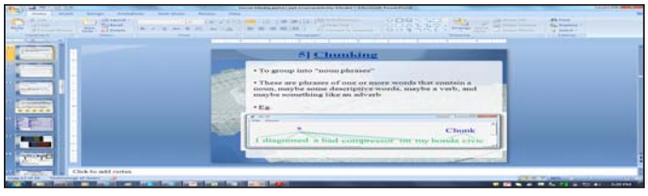
Text Cleansing Process: Majority of available text data is highly unstructured and 'noisy' in nature – to achieve better insights and to build better algorithms, it is necessary to clean textual information from various junks e.g. typos, bad grammar, usage of slang, presence of unwanted content like URLs, stopwords, expressions etc. are the usual suspects.The extracted data was cleaned using Natural Language Processing (NLP) technique in Python which involves 7 major steps viz. sentence & word Tokenizing, Removal of Stopwords, Stemming, Parts of Speech Tagging (POS Tagging), Chunking, Chinking and Named Entity Recognition.

Certain Examples:

Chunking

- To group into "noun phrases"
- These are phrases of one or more words that contain a noun, maybe some descriptive words, maybe a verb, and may be something like an adverb

Example:



Removing Stopwords

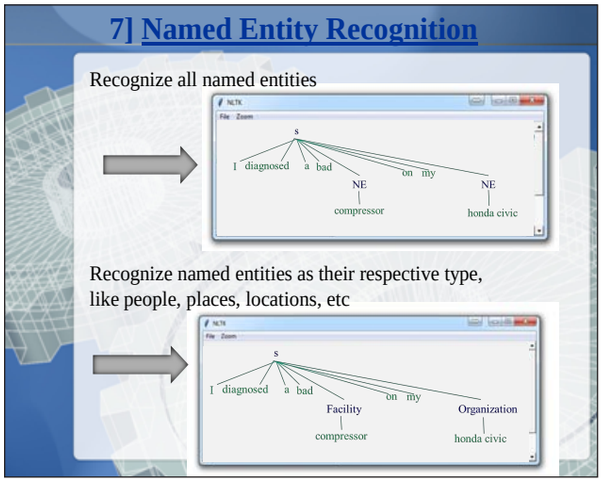
Example:

['T', 'diagnosed', 'a', 'bad', 'compressor', 'on', 'my', 'honda',
'civic', 'yesterday', '.']

Post Stopwords removal:

['diagnosed', 'bad', 'compressor', 'honda', 'civic', 'yesterday', '.']

Named Entity Recognition

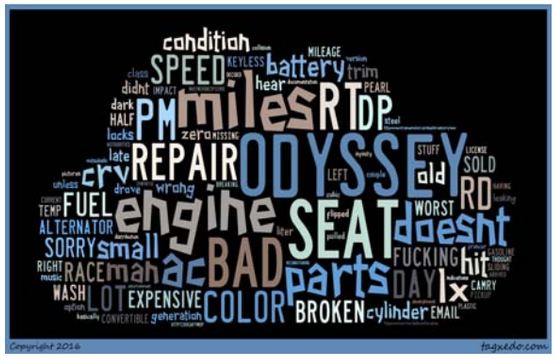


Text Mining: The data integration process is concluded with bringing in disparate variables together to create a tabulated data. Text mining, a procedure of synthesizing information, analysing relations, patterns, and rules among textual data. This included word cloud formation, text summarization, text categorization, and text clustering. Text summarization was used to extract partial content reflection. Text categorization on the other hand, assigned a category to the text amongst the predefined categories. A word cloud is a visual representation of text data used to depict keywords.

Larger the word in the visual the more common the word is in the text data. Segregating the comments as per the polarity (negative and positive). As example, two word clouds are shown below for Honda Brand (for +ve & -ve sentiment) to depict keywords.

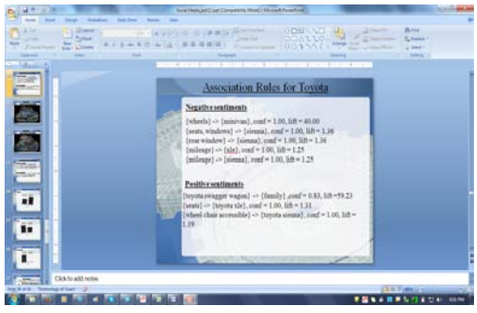
Honda Positive Sentiment

Honda Negative Sentiment



Further with the help of 'Association rule', the content category -sentiment pattern was further strengthened with probability. The rule depict a pattern that states when X (content in text) occurs, Y (sentiment) occurs with certain probability.

Example:



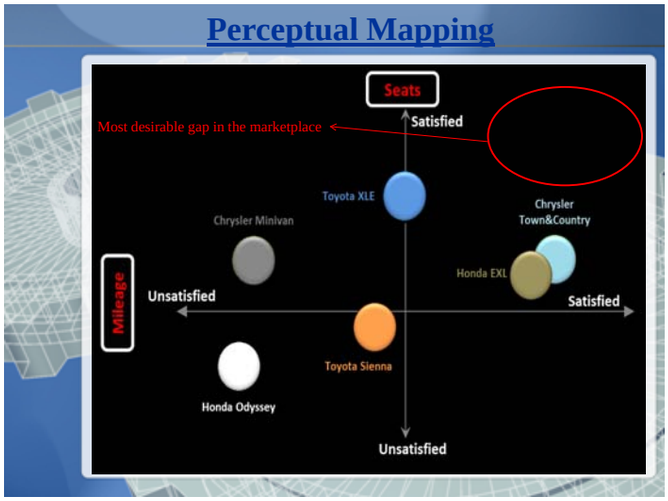
ePercept

Correspondence analysis, or perceptual mapping, is used

to transform numerical information into a graphical display of a particular market or topic area. Relationships among multiple variables (e.g. brand attributes and brands) are represented in two-dimensional space. Various perceptual distances or proximity between brands and attributes can be compared to gain insight into how various brands are positioned against one another in the minds of the target audience, how each attribute compares to other attributes, and where the brands sit in perceptual space.

We used perceptual mapping to show A particular brand performance vis-à-vis its competitor brands which helped marketer's understand where the brand was lacking as compared to competitor brands.

Output Example with Two Brand attributes 'Seats' & 'Milage'



Solution Usage & Impact:

Post development of the solution, periodical delivery with 'Brand imagery' (perceptual mapping) of 'The Study Brand' vis a vis competitors which helps gauging customer's expectations and experience.

With appropriate action, The Brand enjoyed an impact of:

- 4% - 8% Improved CSAT score post actions implemented for each Dimensions/Attributes viz. Purchase, Features, Usage & Service.
- Improved Service Management and Brand Advocacy.

Strategic Action Points:

- Continuous monitoring of brand imagery customer satisfaction, and product feature/service improvement.
- Proactive engagement to enhance customer experience.
- To take focused marketing effort to better position vs. competition and overall market structure.
- To track movement / improvement of themselves and competitive brands.

Article can be accessed online at http://www.publishingindia.com

The Exponential Learning Curve of Analytics

Sudhir Voleti*

We humans, being human after all, are hardwired to think linear. We find linear progressions natural and are able to better understand, explain, predict and evaluate them. Which is why truly exponential changes often tend to stump us. And this is true cutting across domain area, sector and even time. Consider the following famous quote attributed to a famous person:

"We always overestimate the change that will occur in the next two years and underestimate the change that will occur in the next ten." ~ Bill Gates

A big reason for said over- and under-estimation is our simple inability to "get" change along exponential curves the way we "get" linear ones. Let me start with a motivating example.

The date is March 13, 2004 and the venue is California's Mojave desert, site of the DARPA (Defense Advanced Research Projects Agency, of the US govt) Grand challenge in autonomous vehicle technologies. At stake is $1 million in prize money. The task is to design and execute an autonomous vehicle design that will navigate and complete a 150 mile race course having numerous [small] obstacles such as sharp bends, turns, hillocks and mounds, large rocks etc. There were 15 participants. So what happened? Well, none of the vehicles could do even 10% of the course. The relative best performer was CMU's (Carnegie Mellon University) modified Humvee, which did some 7.5 miles before (literally) crashing into a ditch. So DARPA declared the Grand challenge a bust-up and kept the prize money with itself. However, not all was lost. DARPA did manage to see some potential and some promise in the technologies and innovations on offer.

So one and a half years later, on October 8, 2005, at more or less the same venue, it called for a re-match. The prize money was now doubled to $2 million. Also, the obstacles were now tougher – at least three tunnels, narrow roads along cliff-edges, etc. So, what happened this time, a mere 19 odd months after the first (flop) show? Well, this time, among a dozen odd participants who registered, five actually managed to complete this (tougher) race. What's more, four did so within 7.5 hours. The winner was a creation of Stanford's Sebastian Thurn, which beat CMU's entry by merely a 10 minute margin. Oh, it's not over yet. The next date of interest is November 10, 2007, a mere 2 years later. DARPA again calls for a re-rematch, if you will. The prize money's bigger now, too. So also is the level of challenge. Because this time, it is all happening in an urban setting. The rules require this time that the autonomous vehicles obey all of California's traffic laws and demonstrate the ability for such things as merging into traffic, parking by the kerb etc. An astute reader would by this start wondering how they ever got permissions to run a race like that which potentially puts the safety of ordinary human drivers on the road at risk. What actually happened was that DARPA cleared main street and a few other streets of a Californian small town for a day, and hired some 300 professional drivers driving regular vehicles of all types to act as regular American drivers on a regular day out in a small town. OK, so what happened this time? Well, again 5 vehicles completed this, much harder task. On top of the points tally (again) was Dr Sebastian Thurn's creation followed closely by CMU's vehicle. However, later, it was found that Dr Thurn's vehicle had incurred a couple of minor traffic violations (stopping late at a stop sign) and hence was dunked points for the same, pushing CMU to top spot. OK, so what happened next? Not hard to guess by now, I guess. A year later, in 2008, Google launches its self-driving car program with who else but Dr Thurn as its head.

So what was the point of this example? One point certainly is to demonstrate, though an easy-to-see example the phenomenal advances in configuring an

* Indian School of Business, Hyderabad, Telangana, India. Email: Sudhir_Voleti@isb.edu

entire battery of technologies -think of sensors (to keep tabs of the vehicle's internals as well as immediate external environment), radars (to assess relative speeds in own and neighboring vehicles and thereby collision risk), basic machine vision (to spot and read traffic lights, stop signs and pedestrians among other things), etc. These advances kept pace with a set of task challenges whose difficulty level rose exponentially. Two, the design and engineering teams didn't have a whole lot of data to start with. They quickly learned from the first race's failure enough to ensure completion of the next, much tougher track. Thus, the researchers were able to train the machine (the controlling computer(s) in the autonomous vehicles) fairly quickly with limited training data. Three, and this is where things mesh into where I was hoping to go to with this article, it points to how much and how fast the growing cognitive capabilities of machines are likely to scale up, in the coming years.

From IBM's Deep blue beating Garry Kasparov (who some call the greatest human chess player ever) in the structured, rule bound game of chess using brute-force number crucnching capabilities in 1997 to 2011 when IBM's Watson beat the best human players of a much more loosely structured (and pun filled) game 'Jeopardy', the cognitive capabilities of super machines has been on a relentless rise. Very recently,Alphabet's artificial intelligence (AI) project best the world's best 'Go' player. What has now entered this mix are factors such as (a) a flood of data pouring in (on which machines can be trained), and (b) open source software platforms that allow coders, developers, programmers, modelers to collaborate, extend, test, tweak and animate an ever expanding set of algorithms, routines and programs.Thus, someone on one side of the planet working on a different explanation, prediction or optimization problem could benefit from a fix someone on the other side of the planet came up with for a different problem in perhaps a totally different area. What results is borrowing, tweaking, extending, bug-fixing, bridging (with other computational platforms) and other such processes happening 24x7 in the open-source world. I often visualize open source routines and packages coming up as a set of Lego blocks adding to a growing repository, which anyone could configure in a way that builds something very neat, useful and perhaps unique. Where this potentially leads to is the possibility of advances across different fields bootstrapping one another and spurring on (in bursts, fits and starts) exponential learning curves in certain domains.

Notice that the motivating example focuses on things that move in the physical world ('atoms') as opposed to just in the virtual world ('bits'), though the case for the latter in terms of exponential learning curves remains just as strong. Consider ROS, or 'Robot Operating System' built on the same open-source principles that currently drive progress in an R or a Python. The race is on to enhance-capability, cheaply, of basic robots using programming to make them smarter and optimize what physical world movements are currently allowed them by their design, sensor and tooling configurations. When Microsoft launched the Kinect, which became a runaway hit, what happened in the robotics world afterwards is worthy of note. Within weeks of Kinect's release,it had been hacked for machine vision applications by enthusiasts all over the world and the videos were posted on youtube. And why not? You now had substantial quality machine-learning capabilities built into a widely available set of hardware and sensors and priced at a mere few hundred dollars (relative to what machine vision applications would cost in the pre-Kinect era). Now consider a similar story playing out in advances in other fields not just sensors such as software applications. Consider what happened the last time a standard OS + inexpensive programming tools became available? What ROS is trying to do is to leverage open source architectures to build exponential learning curves into Robotics development, potentially threatening to disrupt entire industries and sectors along the way.

To quote an old Chinese curse, we do live in interesting times. Fasten your seatbelts and hold on tight. It'll be quite a ride ahead.

Article can be accessed online at http://www.publishingindia.com

Airline Revenue Management – Revenue Maximization Through Corporate Channel

Karthik V*, Indranil Mitra*

Abstract

This paper explains an empirical model that has been developed to arrive at the Best Possible Fare (BPF) for all the ad-hoc requests made by the corporate passengers. For the purpose of this research, a corporate request is considered ad-hoc if the booking request is initiated after the corporate channel for the particular flight is closed. By charging the best possible fare, the airline will be able to marginally increase its revenue without deviating from the guidelines of the corporate channel. This model updates itself with the available capacity at the time when the ad-hoc request is initiated, also considers the previous booking data to forecast the passenger demand and the channel behavior. This will lessen the manual intervention and its associated errors, and will take care of the number of corporate requests that can be approved and size of the corporate booking requests that can be approved. As the factors affecting the booking trend of the airlines have been covered earlier in various research papers as discussed in the literature review, we have directly focused on deriving the empirical solution in this paper.

Keywords: Revenue Management, Best Possible Fare, Forecasting, Airlines, Corporate Passengers, Sales Channel

INTRODUCTION

After the deregulation of airline industry in India, all the airline organizations started offering differential price for the seats in the same flight and in the same cabin. Different airlines use different pricing rules and algorithms to arrive at the final price for each seat which is to be sold at a future date. With the invasion of different platforms in the form of travel agents and online booking systems to compare and book tickets, the Indian Airline Industry is suffering to maximize its revenue in the low cost game. This part of how many seats should be offered at what rate and for which period of time before booking for a flight closes, is taken care by revenue management team. The importance of revenue management to dynamically vary their fares according to the supply and demand is essential in the path towards maximizing the overall revenue. Each and every airline manages to fulfill this objective depending on the level of sophistication of the algorithms and scientific techniques incorporated into the system.

Typically, historic booking data, information of seats sold through different channels and its forecasts are used to predict the future demand. This forecast serves as an input for the revenue optimization step, which considers capacity and fares (Bamberger, G.E., D.W. Carlton and L.R. Neumann., 2004). The resulting inventory controls are balanced with real world demand when taking reservations. Demand is strongly influenced by market conditions. The results obtained are again used as input data for the next demand forecast and so on.

In airline industry the customers can be broadly divided into categories; corporate and leisure (Kelly Mc. Guire, 2012) based on their ticket booking behavior. As the need to travel on a particular date is different for each of these two categories, the airline uses differential pricing methodology to extract maximum value out of the customer to enhance its revenues. The price of the ticket for a leisure passenger increases as the booking date comes closer to the departure date. This is because, these passengers who look for tickets during that time have higher propensity to travel and hence will be ready to pay the premium to purchase the ticket. As these customers are individual customers and they are not bound by any contract, the airline does not see any benefit in giving them a discount on the maximum value that can be extracted from the customer. However the price of the ticket for corporate passenger does not vary much because of the volumes of the business the corporate brings to the airlines.

* KPMG-India, D&A-Management Consulting Gurgaon, Haryana, India.
E-mail:karthikv606@gmail.com, Indranil.mitra@gmail.com

In the case of corporate passengers, the airline enters into a contract with the corporate customer to provide special benefits with less degree of variation in the ticket price without any effect on the date of travel. Due to this reason the airline is not able to alter its fare in case of a special request from an existing corporate customer leading to opportunity loss in terms of selling tickets at increased price through other channels.

The change in the pricing structure is governed by various pricing rules and algorithms written for the same. It is observed that in general the pricing algorithm is a factor of demand, competition pressure, sales channel, and booking time as a function of days before departure among others (Fedorco & Hospodka, 2013).

This paper is laid out as follows: In section 2, we have illustrated the current scenario as the problem definition and emphasized on the scope for improvement in the same section. To build an objective case we have also made a few assumptions which are also mentioned in the same section. In section 3, we have reviewed the contribution of a few research papers in exploring how dynamic pricing across different channels has been handled in the past. In section 4, we have explained about our model and the approach considered in building the formula. In section 5, we have developed the empirical formula that will give us the best possible fare that should be charged to accept the special request from the corporate customer. In section 6, we have discussed the results and consolidated the investigations. In section 7, conclusion of the paper is presented.

PROBLEM DEFINITION

In Industry various revenue management solutions are available in the market to help airlines arrive at the optimal passenger mix for a scheduled aircraft. But these solutions are built to recommend the optimal passenger mix of corporate and leisure passengers at a suitable time before the seats are booked for a particular flight. Based on the recommended passenger mix, the airline opens its inventory for both classes of passengers across various channels. Once the aircraft reaches its recommended passenger mix for a particular type of passenger, the inventory for that particular type of passenger is blocked and the remaining seats are allotted for the other passengers.

Consider this scenario: The corporate bookings for a particular flight with departure date 7 days from now are closed, as the revenue through corporate channel has reached the desired limit. Now there is a special request of size 'S' for additional bookings from one of the corporates for the same flight. The revenue management team has to decide whether to accept the request or not.

Under the current scenario, the airline

1. Accepts the request, if it can achieve the budgeted flight revenue.
2. Accepts the request, if the corporate is a loyal customer.
3. Rejects the request otherwise.

In each of these conditions, the decision is arrived at by performing basic projection of the commercials based on past performance and future relationships. The airline does not use any scientific reasoning to forecast the opportunities across other existing channels. A standard revenue management solution does not have the provision to accommodate the ad-hoc requests. The requests, if approved, are approved at the previously determined rates. Hence the airline is not able to leverage the left out inventory to make additional revenues through more profitable channels.

To approve 'S' seats for corporate ad-hoc requests, the airline displaces 'S' seats from some other channel. Since the price of seats in other channels are highly sensitive to the date of travel, while borrowing seats from other channels the airlines must consider the opportunity cost and arrive at a suitable price to avoid losses. The corporate ad-hoc request can be accepted as long as the revenue generated from this scenario is at least equal to the expected revenue that the corporate passenger displaces.

In the scope of the paper we do not consider the external factors that impact the booking trend before the departure of the flight. External factors like natural calamity, any unprecedented event in the arrival or departure destination and others. We have predominantly focused on the industry trend and controllable seasonal factors, which under normal scenario do not alter the booking trend of a flight.

Assumptions: To make this analysis quantifiable, we have made the following assumptions in the paper.

1. When a passenger considers to travel, he/she approaches the airline mentioned in the corporate agreement.

2. We have considered corporate passengers only through corporate channel as it difficult for us to differentiate between passengers through other channels.
3. We have considered customers as if purchasing one way tickets only. Since there is less difference between the price of one way and round trip tickets a customer who has purchased a round trip ticket is considered to have purchased two one way tickets.
4. We have assumed airlines utilize finite set of fares. Though the airline sells the inventory at different prices across different channels, airlines chooses a single set of fare to offer at a particular period of time.
5. Cost centers are fixed. Since the marginal increase in the revenue does not come at the cost of any operational expenses, increase in revenue leads to increase in profitability.
6. Finally, we assumed that airlines do not oversell tickets as most of the routes studied are operated by airline that do not oversell tickets.

LITERATURE REVIEW

The question of identifying the optimum passenger mix and setting the price for each channel and time period has been discussed a lot in the academic literature.The considerations that started much of revenue management research can be found in Littlewood (1972). He mathematically formulated an intuitive rule that proposes to sell tickets in the cheaper of two booking classes, as long as the expected marginal utility exceeds the fare of the more expensive booking class. Later, Belobaba (1987) extended this approach to more than two booking classes, resulting in a still frequently applied concept: expected marginal seat revenue (EMSR). A review of developments and future directions of revenue management are given by McGill and van Ryzin (1999). Furthermore, Talluri and van Ryzin (2004) provide a comprehensive insight into the concept of revenue management and its elements. A more recent overview of mathematical models and methods in revenue management and its focus on simulations can be found in Talluri et al. (2008).

There are two important streams of research on revenue management: (i) empirical studies on airline pricing methodologies, and (ii) analytical revenue management models in the literature. In the economics literatures, studies have empirically examined the relationship between airline pricing and various market factors. Borenstein and Rose (1994) find a significant positive effect of competition on price variations in the airline industry. Hayes and Ross (1998) find airlines' price discrimination policies lead to increased price variations. Borenstein (1989, 1990) finds that airport dominance enhances a carrier's ability toattract passengers and charge higher fares. This may be attributed to biases due to computerreservation systems, the dominant carrier's local reputation, control of critical inputs such asgates and slots, and marketing strategies such as frequent flier plans (Evans and Kessides 1993).Peteraf and Reed (1994) find that a monopolist's national market share has a positive effect onfares and that prices tend to decrease in the number of passengers and route distance.

Moreover, they find that the average code-share fare is lower than the average fare that isnot code-shared. Bamberger et al. (2004) also find that the price tends to decrease after alliances.Their findings are similar to those of Park and Zhang (2000), Brueckner and Whalen (2000), andBrueckner (2001, 2003) who examined international alliances.

The quantity-based revenue management models start with Littlewood's seminal work (Littlewood 1972, henceforth referred to as the Littlewood model). The Littlewood model studies how the fixedtotal capacity should be allocated between two classes of seats once fares are determined. The model assumes a fixed number of seats and two independent classes of demand-demand forfull-fare tickets and demand for discount-fare tickets. Discount-fare demand occurs first, and it is large enough to fill all the allocated seats. The demand for full-fare tickets occurs later and is random. The model derives the optimal seat protection level for full-fare demand. The analysis of the problem is similar to that of the classical newsvendor problem in the inventory theory (Talluri and van Ryzin 2004). The Littlewood model has since been extended to multiple-classmodels (Belobaba 1989, Wollmer 1989, Curry 1990, Brumelle and McGill 1993, Robinson 1995) and dynamic models (Lee and Hersh 1993, Feng and Xiao 2001).

For price-based revenue management models, the seminal work of Gallego and van Ryzin (1994) analyzes the optimal dynamic pricing policy for one type of product. Gallego and van Ryzin's dynamic pricing model assumes that consumers arrive randomly. The optimal price has the following important properties: (i) At any fixed point intime, the optimal price decreases in the inventory level; conversely, for a given level of inventory level, the optimal price increases with more time to sell. (ii) For a

fixed time and inventory level,the optimal price increases in the arrival rate. Zhao and Zheng (2000) extended this model to thecase where demand is non-homogeneous. Since consumers are time sensitive, their reservation price distribution may change over time. For a good review of the current practices in dynamic pricing, see Elmaghra by and Keskinocak (2003).

The Littlewood and GVR models offer important insights that will be used in our discussion of the empirical findings. Shumsky (2006) finds that low-cost competitors are driving the network airlines to rely on alliances for an increasing proportion of their traffic. Wright et al. (2006) study a variety of static and dynamic mechanisms to manage revenue management decisions across alliances.

Using these mechanisms the airline is able to change its pricing for the same ticket based on the channel and the time period of search of the ticket. Airlines tend to charge high prices from passengers who search for tickets close to the date of travel. The conventional view is that the airlines capture their high willingness to pay through inter-temporal price discrimination. Airlines also adjust price on a day-to-day basis as capacity is limited and the future demand for any given flight is uncertain. While fares for a leisure passenger generally increases as the departure date approaches, fares for the business passengers are not altered much because of the corporate agreements and the quantum of business it generates during the engagement. By keeping the ticket price constant irrespective of the time period of ticket being booked and the type of corporate the airlines are losing out on the opportunity to marginally increase the revenue. Hence in this research paper, we have considered the factors impacting the sale of tickets through corporate channel and arrived at the best possible fare that should be charged from a corporate to maximize the revenue without breaching the guidelines of the channel.

Our Work

Based on the ticket booking behavior all the airlines have divided their target customers into leisure and corporate passengers at a broad level as mentioned in the literature review. Each airline has its own strategy to tap its target audience i.e. channel, pricing and services offered for the same seat in a particular flight.

Most of the airlines today manage corporate bookings through their specific corporate booking channel and also through specialized indirect agents who help in acquiring only corporate customers. The entire process of acquiring a corporate customer is bound by a contract between the airline and the corporate customer, hence there is less manual intervention at the time of booking the ticket. This leads to less flexibility in associating the latest available price of the ticket to the corporate customers. Since major part of the corporate bookings happen less than one week before the date of travel, the price of a ticket during that time through corporate channel is less than the cost of the same ticket through other non-corporate channels.

In the case of a flight where the corporate bookings are closed after the channel reaches its desired load factor any additional request is separately handled by the revenue management team. The entire process involves manual intervention from the airline side – right from enquiry for additional corporate booking through approval, booking, modifications to the itinerary, payment and ticketing. Entire process is resource intensive, time consuming and error prone right from enquiry to final billing. Also if the corporate special request gets approved the ticket is booked at a predefined pricing rule as per the contract. It is generally observed that the same ticket through other channels is higher than what is actually sold. Hence the airline has lost the opportunity to earn additional revenue by selling the ticket at a lower price through corporate channel. The difference in the pricing through these channels is so high that the airline ends up selling the ticket at a price less than what it could have sold by charging a premium without comprising the benefits offered to corporate passengers for the bookings done during the same time. The problem is further complicated by last minute cancellations and modifications on PNR resulting in inventory being blocked which otherwise can be sold through other channels and then maximize the revenue quotient for the airlines.

Keeping the above factors in mind we have arrived at an empirical solution from a representational sample of three months booking data of a leading airline organization. This empirical solution is helpful in providing decision support through scientific forecasting and optimization techniques to arrive at the best possible fare that needs to be charged to accept the corporate special request below which it does not make business sense to accept the corporate special request. This will lead to marginal increase in revenue by charging a premium to approve their special requests without deviating from the agreement. This is

done with respect to analyzing the historical booking data at channel levels, corporate booking activities, utilization rates of corporate bookings and many more.

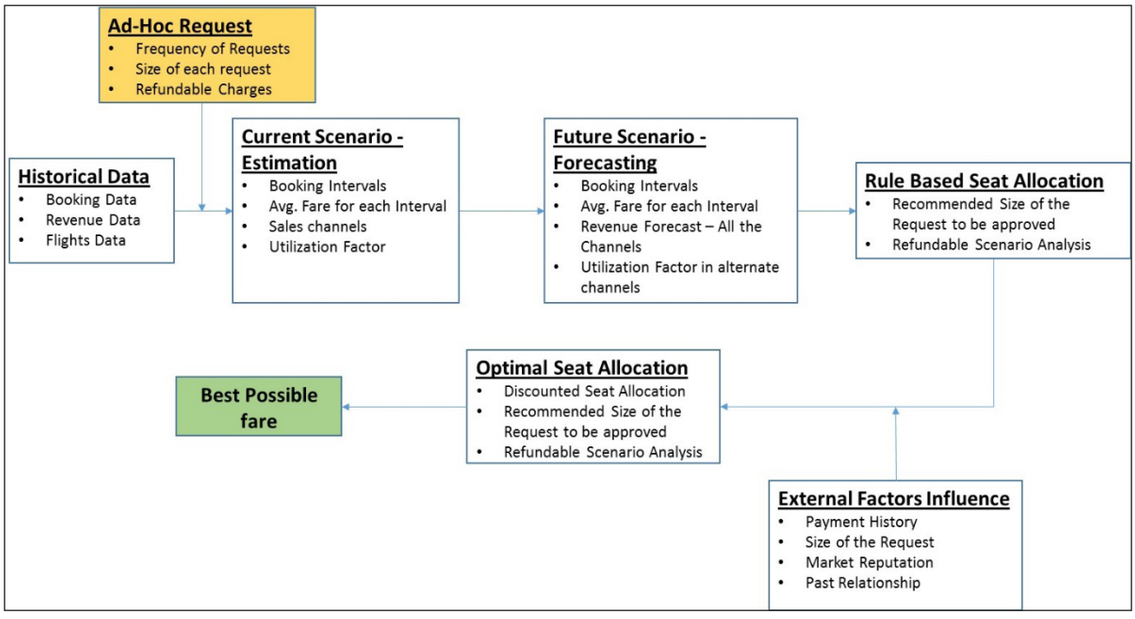
METHODOLOGY AND FRAMEWORK:

Most airlines today manage their corporate booking through their specific corporate tie ups in the form of direct agreements with a corporate or indirectly through specialized corporate agents who are directly associated with the corporate. The decision of accepting the request, pricing of tickets, ensuring the arrived price is the maximum value that can be charged for the booking without losing the customer, special requests like internal travel can be complex based on the requests and cancellations. The entire process involves manual intervention from the airline revenue department rendering the whole process tedious and error prone.

The figure below gives a schematic layout of the approach adopted to calculate the best possible fare.

This methodology

1. Provides an optimal airline seat allocation scheme for the revenue management team for their additional corporate requests
   a. Number of seats for corporate booking process
   b. Number of corporate bookings request
   c. Size of each booking request
   d. Corporate Price Per Seat
   e. Revenue Forecast
   f. Refunds in case of cancellation
2. Provides an optimal control mechanism for the corporate booking process
   a. Demand Pattern
   b. Booking Pattern
   c. Cancellation Pattern
   d. "No Show" Pattern'
   e. Overbooking
   f. Corporate booking Start Date
   g. Corporate Booking End Date



   h. Time between negotiation phase and placement of non-refundable deposits
   i. Time between non-refundable deposits and actual price of tickets
   j. Time between actual purchase of tickets and date of departure

Thus, there are several other factors that need to be incorporated in a structured manner to come to a reasonably good solution.

A random sample of the past three month's data was extracted for all the considered routes (metro). A total of 35 variables were chosen, of which through correlation analysis, 9 variables with higher degree of correlation

were dropped. Remaining 26 variables are considered for the study. Subsequently we performed ANOVA test to identify the degree of significance of each variable. Since 8 variables had p>0.05 we have finally kept only 18 variables for our study. The Cronbach alpha of the chosen set of variables is found to be 0.6723 which is within acceptable limits for the study. The list of the variables used for ANOVA test and their p value are presented in Appendix 1.

To consider only flights where the number of ad-hoc requests received are consistently more than one we have considered all the flights with load factor > 80%. In terms of number of flights this data forms 40% of the entire dataset considered.

To arrive at the best possible fare we need to calculate the Total Expected Corporate (TECR) revenue from the corporate special request to make the flight profitable. TECR is calculated by analyzing all the possible scenarios of seat allocation by taking booking intervals, utilization factor for each channel among others into consideration. The Next step is to calculate the Total Expected Flight Revenue (TER) for the same number of seats displaced from other channels. In this step, we will be using statistical forecasting techniques to calculate the forecasted revenue across all the channels. If the difference between TECR and TER to fill the same number of displaced seats under normal scenario is positive then calculate the best possible fare by dividing the TECR by size of the request. If the obtained average fare is less than the average fare as mentioned in the agreement then the revenue team quotes the price as mentioned in the agreement.

MODEL

Corporate booking process starts about 12 months prior to the flight departure. Let the time span from the start of the booking till flight date be divided into suitable intervals as T = t, t-1, t-2,…, 1, 0. Where t=0 denotes the actual flight departure time. For each time point we have the available seats, demand for the available seats across each channel, average fare for the seats and the actual request that comes for the corporate booking. To illustrate further consider the following:

Let,

The set be defined as:

I – Class of fares for an airline

K – Total Time Interval starting from the start of corporate booking till departure of flight

Let the variables be defined as,

At each point there are 3 components to corporate booking process:

D_{ij} – Demand that is estimated from the historical data at the time point 'j' for the 'I' class

F_{ij} – Average Fare that is to be estimated from the historical data at time point 'j' for class 'I'

μ_{ijr} – the corporate request that comes at the time point 'j' for the class 'I'

C_i – the number of reserved seats for corporate booking for the class 'I'

G – Total number of booking class

$$\sum_{i=1}^{g} Ci = C,$$

F_{ik} – Average Fare for the i^{th} class at a particular time interval 'K'

C – Corporate booking limit for the i^{th} class

– Opportunity cost of the airlines because of not selling the seats through some other class of service

– Fraction of originally booked tickets at k^{th} point of time.

– Opportunity cost of the airlines because of passengers who paid for the seat but did not show up

– Fraction of originally booked tickets at k^{th} point of time.

Decision Variable

– The number of seats that will be allotted to a corporate request (r) of size at time point 'k'.

Here we are assuming that a corporate request of μ at a particular time point comes for the i^{th} class.

Mathematical Model

The base model that would be used for optimization can be given as:

$$z = Max \sum_{r} \sum_{k} \{ (\sum_{i=1}^{g} F_{ik} X_{ik}) - \left(C_{1kr} * P_{ikr}^{c} + C_{2kr} * P_{ikr}^{c} \right) \} \square(P) [$$

Subject to

$X_{ikr} <= \min(d_{ij}, \mu_{ikr})$ (1)

$\sum X_{jkr} <= C$ (2)

At any point 'K', gives the revenue earned over all the classes for a particular flight. On the other hand $(C_{1k} * P^{C}_{1k} + C_{2k} * P^{C}_{2k})$ gives the total opportunity cost that the airline incurs at the time point 'K' taking into consideration the different situations as described before. Thus subtracting the opportunity cost from the revenue earned gives net revenue earned at time point 'K'.

Constraint (1) implies that the number of seats offered would be less than or equal to the number for which order is placed.

Constraint (2) implies that the number of seats offered in all classes is to exceed the capacity offered in corporate request.

Model output of P gives the size of the corporate. Solving the objective function of (P) gives the maximum revenue that can be earned at k^{th} point of time where k = K.

Best Possible Fare Calculation

In accepting a corporate request of size 'S', an airline potentially displaces up to 'S' individual passengers. Since corporate fares are discounted below the fares through other channels as the booking date gets closer to the departure date, the decision whether or not to accept a corporate request depends on individual passenger on each flight flying in comparison with corporate passenger. This corporate request should be accepted as long as it makes business sense to the airlines. This is termed as the total expected revenue of the displaced passengers(TERDP).

Best Possible Fare (BPF) = TERDP/No. of approved Corporate Request

Where TERDP is calculated from the historic booking trend of the same aircraft for a similar booking season.

Let us define Z(C) to be the optimal objective value function (I) using the initial set of corporate booking limit C.

Now consider an ad-hoc request of size 'S'.

Z(C-S) – is the optimal objective function solving (I) where the capacity constraints of each aircraft where the corporate passenger will travel is decreased by the size request S.

The value Z(C-S) is the best solution of the problem given that one has accepted the corporate request and S seats are no longer available for further passenger booking.

We define the difference of the objective functions Z(C) and Z(C-S) to be D(S) which is defined as TERDP. Thus D(S) represents total expected revenue of displaced passengers.

In algorithmic form the following can be proposed:

Step 1: Find Z(C) using the linear mathematical programming formulation (P) for the given network.

Step 2: Reformulate the mathematical program to calculate Z(C-S), where the capacity constraints used in step 1 reduced by S where the group travels. The reformulated model isgiven as:

$$Z = Max ?_{(i=1)}^{g} ?? (F_i k X_i k) - (C_1 k P_1 k^c + C_2 k P_2 k^c \} \text{ (Q)}$$

Subject to

$X_{ik} <= \mu_{ik}$ (1)

$\sum X_i + S <= C$ (2)

k = K, i = I;

Note here that the constraint (2) states that the capacity of the seats that has been reduced by 'S'

Step 3: Find D(S) = Z(C) – Z(C-S)

Step 4: MAF = D(S)/S

The steps are to be executed in a looping procedure where the next starting point would beZ(C-S). The procedure would continue till all the seats are used up or group request period ceases, whichever may be earlier.

RESULTS

Airlines manage corporate ad-hoc requests based on the potential of business from the corporate. The entire process involves manual intervention from airline revenue department rendering the whole process tedious and error prone.

Using this empirical solution the airline will be able to know the best possible fare below which, by approving

the corporate request the airline is losing the opportunity to earn additional revenue. Making an informed decision will also help the airlines plan its future inventory across various channels and adjust its targets to achieve overall profitability. We have summarized below the results that can be obtained by using this scientific way of addressing the corporate ad-hoc request:

S.No.	Factors	Description	Drivers
1	Determine the Best Possible Fare(BPF)	BPF for a given O&D, Date	Displacement Cost Calculation, Revenue Optimization and Inventory Management
		Extend BPF calculations to multi leg journey	
		Self-adjustable BPF due to nearing departure time, cancellations and increased demand	
2	Calculate Discount Rates to be given	Determine discount percentage to be applied on fares based on past performance	Frequent Flyers, volumes of business by the corporate
3	Alternative Travel Plan	Requests that cannot be accommodated for can be re-routed for an alternate date	Network Plan along with Booking Details, Displacement Cost

CONCLUSION

We have used the representative data of a leading Indian airline to examine the standard revenue management practices and perform our analysis. We have concentrated on the revenue opportunities through corporate channel. Because our model includes market factors affecting the revenue of the channel, our empirical formula can also be extended to model other channels.

This paper explains the empirical model to arrive at the best possible fare, required to increase the revenue and hence the profitability without altering the benefits offered to the corporate customers. This will help the airlines to marginally increase their revenue for the same services offered to the customer based on the time of the request, size of the request and passenger mix of the particular aircraft. This analysis is not helpful in a case where the airline feels it is necessary to accept the request as a directive from a higher official without thinking about the impact of accepting the request on its revenue and profitability. Since the number of such important requests are going to be very less as such decisions will have a negative impact on revenue, we have ignored such scenario from our analysis.

Our model helps the revenue management team to arrive at the best possible fare for the corporate ad-hoc request. The current work has the limitation that, in case there are empty seats closer to the date of departure, the model can be extended to come up with discounted fares for high performing corporates as a reward for their loyalty.

References

Alderighi, M., Nicolini, M., & Piga, C. (2012). *Combined effects of load factors and booking time on fares: Insights from the yield management of a low-cost airline.* Milan, Italy: Nota di Lavoro, Fondazione Eni Enrico Mattei.

Bamberger, G. E., Carlton, D.W., & Neumann, L. R. (2004). An empirical investigation of the Competitive effects of domestic airline alliances. *Journal of Law and Economics*, 47, 195-222.

Baltagi, B., Bresson, G., & Pirotte, A. (2003). Fixed effects, random effects or Hausman-Taylor?A pretest-estimator. *Economics Letters*, 79(3) 361-369.

Belobaba, P. (1989). Application of a probabilistic decision model to airline seat inventory control. *Operations Research,* 37, 183-197.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity.* John Wiley & Sons, New York, N.Y.

Bergstrom,T., & MacKie-Mason, J. (1991). Some simple analytics of peak-load pricing. *Rand Journal of Economics,* 22, 241-249.

Borenstein, S. (1989). Hubs and High Fares: Dominance and Market Power in the U.S. *Rand Journal of Economics*, 20(3), 344-365.

Borenstein, S. (1991). The Dominant-firm advantage in Multiproduct Industries: Evidence fromthe U.S. Airlines. *The Quarterly Journal of Economics,*106(4), 1237-1266.

Borenstein, S., & Rose, N. L. (1994). Competition and Price Dispersion in the US Airline Industry. *The Journal of Political Economy*, 102(4), 653-683.

Brueckner, J., & Whalen, T. (2000). The price effects of international airline alliances. *Journal ofLaw and Economics,*43, 503-545.

Brumelle, S. L., & McGill, J. I. (1993). Airline seat allocation with multiple nested classes.*Operations Research*41, 127-137.

Burdett, K., & Judd, K. L. (1983). Equilibrium price dispersion. *Econometrica*, 51(4), 955-969.

Cachon, G., & Terwiesch, C. (2006). *Matching supply with demand.* McGraw-Hill/Irwin.

Clark, R., & Vincent, N. (2012).Capacity-contingent pricing and competition in the airline industry. *Journal of Air Transport Management*, 24, 7-11.

Carpenter, D. (2007). *Fuller planes pay off for UAL bottom line.* The Grand Rapids Press, July 25.

Curry, R. E. (1990). Optimum seat allocation with fare classes nested by origins and destinations. *Transportation Science,* 24, 193-203.

Dresner, M., & Windle, R. (1992). Airport dominance and yields in the U.S. airline industry. *Logistics and Transportation Review*, 28(4), 319-339.

Eblen, T. (1996). The grounding of Value jet Airlines must try to fill all seats, charge the highest possible fare. *The Atlanta Journal-Constitution,* June 23.

Elmaghraby, W., & Keskinocak, P. (2003). Dynamic pricing in the presence of inventory considerations: research overview, current practices, and future directions. *Management Science,* 49(10), 1287-1309.

Evans, W., & Kessides, I. (1993). Localized market power in the U.S. airline industry. *The Review of Economics and Statistics,* 75, 66-75.

Gallego, G., & van Ryzin, G. (1994). Optimal dynamic pricing of inventories with stochasticdemand over finite horizon. *Management Science,*40, 999-1020.

Greene, W.H. (2003). Econometric Analysis.Prentice Hall, Upper Saddle River, N.J.Hayes, K. and L.B. Ross. 1998. Is price dispersion the result of careful planning or competitive forces? *Review of Industrial Organization,*13, 523-541.

Ito, H., & Lee, D. (2007). Domestic codesharing, alliances and airfares in the US airline industry. *Forthcoming in Journal of Law and Economics.*

Williams, K. R. (2013). *Dynamic airline pricing and seat availability.* Department of Economics, University of Minnesota (Working Paper).

KIMES, S. E. (1997). *Yield Management: An Overview.* Yield Management, Strategies for the Service Industries. London: Routledge.

Kimes, S. E. (1989). Yield management: a tool for capacity-constrained service firm. *Journal ofOperations Management* 8, 348-363.

Lee, T. C., & Hersh, M. (1993). A model for dynamic airline seat inventory control with multipleseat bookings. *Transportation Science,* 27, 252-265.

Littlewood, K. (1972). Forecasting and control of passengers. *12th AGIFORS Symposium Proceedings. Nathanya, Israel,* (pp.95-128).

Fedorco, L., & Hospodka, J. (2013). *Airline Pricing Strategies in European Airline Market*, 7(2).

McGill, J., & van Ryzin, G. (1999). Revenue management: Research Overview and Prospects. *Transportation Science,* 33, 233-256.

Netessine, S., & Shumsky, R. (2004). Revenue management games: Horizontal and vertical competition. *Management Science,* 51(5) 813-831.

Peteraf, M. A., & Reed, R. (1994). Pricing and performance in monopoly airline markets. *Journal of Law and Economics*, 37(1), 193-213.

Robinson, L. W. (1995). Optimal and approximate control policies for airline booking with sequential nonmonotonic fare classes. *Operations Research,* 43, 252-263.

Rothman, A. (2006) Sep. 1. Air France-KLM raises profit forecast for 2006 MARKETPLACE by Bloomberg. *International Herald Tribune,* page 13.

Shumsky, R. (2006). The southwest effect, airline alliances, and revenue management. *Journal of Revenue and Pricing Management*, 5(1), 83-89.

Talluri, K., & x van Ryzin, M. (2004). *The theory and practice of revenue management.* Kluwer Academic Puslishers.

Vulcano, G., van Ryzin, G., & Chaar, W. (2010). Choice-based revenue management: An empirical study of estimation and optimization. *Manufacturing & Service Operations Management*, 12 (3), 371-392.

Windle, R., & Dresner, M. (1993). Competition at "Duopoly" Airline Hubs in the US.

*Transportation Journal,*33(2), 22-30.

Wollmer, R. D. (1989). An airline seat management model for a single flight leg route when lower fare classes book first. *Operations Research,* 40, 26-37.

Wright, C. P., Groenevelt, H., & Shumsky, R. (2006). Dynamic Revenue Management in Airline

Alliances, working paper.

Zhao, W., & Zheng, Y. (2000). Optimal dynamic pricing for perishable assets with non homogeneous demand. *Management Science,* 46, 375-388.

Appendix 1

ANOVA test results (main effect) of Independent variables on airline ticket price between two destinations.

S.No.	Factor	Description	F Value	P Level
1	PAX	Number of Passengers	9.289	0.0105
2	Dep Station	Departure Station	6.432	0.0400
3	Arr Station	Arrival Station	1.560	0.0827(NS)
4	Days to Departure	Difference between Booking Date and Departure Date	7.524	0.0295
5	Booking Type	Booking Channel	5.555	0.0474
6	Fare Basis	Fare Basis	3.786	0.0505
7	Class of Service	Code for Class of the ticket	6.885	0.0360
8	ASK	Available Seat Kilometer	0.500	0.2466(NS)
9	Comp_Scenario	Number of competitors Operating between the same stations	9.694	0.0077
10	RPK	Revenue Passenger Kilometer	6.831	0.0360
11	No. of Flights	Number of available flights on the same day	8.467	0.0248
12	Promo Code	Code for Promotions in the corresponding Channel	3.544	0.0621
13	Base Price	Base fare	7.303	0.0343
14	Convenience Fee	Convenience Fee	1.694	0.0790(NS)
15	UDF	User Development Fee	8.972	0.0222
16	AAT	Airport Authority Tax	7.898	0.0283
17	PSF	Passenger Service Fee	0.552	0.1708(NS)
18	GST	Government Service Tax	0.148	0.3734(NS)
19	CUTE Fees	Transaction Fees	1.103	0.0892(NS)
20	Fuel Charge	Fuel Charges	0.684	0.0898(NS)
21	Capacity	Total available Capacity	8.120	0.0248
22	Currency Code	Currency Code	7.495	0.0299
23	STD	Scheduled Time of Departure	5.586	0.0461
24	STA	Scheduled Time of Arrival	9.511	0.0104
25	Flight Tail No.	Registration Number of the Flight	3.876	0.0504
26	L.F	Load factor	2.423	0.0754

NS – Not Significant

Article can be accessed online at http://www.publishingindia.com

Forecasting The Stock Market Values Using Hidden Markov Model

R. Sasikumar*, A. Sheik Abdullah*

Abstract

The financial market influences personal corporate financial lives and the economic health of a country. Price change of stock market is not a completely random model. The pattern of financial market has been observed by some economists, statisticians and computer scientists. This paper gives a detailed idea about the sequence and state prediction of stock market using Hidden Markov Model and also making inferences regarding stock market trend. The one day difference in close value of stock market value has been used for some period and the corresponding transition probability matrix and emission probability matrix are obtained. Seven optimal hidden states and three sequences are generated using MATLAB and then compared.

Keywords: Hidden Markov Model, Transition Probability Matrix, Emission Probability Matrix, Stock Market, States and Sequence

Introduction

The most of the trading in Indian stock market is classified in two categories, the Bombay Stock Exchange (BSE) and the National Stock Exchange (NSE). The BSE has been functioning since 1875. The NSE was founded in 1992 and started trading in 1994. Even though both exchanges have the same trading mechanism, trading hours, settlement process, etc., they are having high demand from people. The two prominent Indian market indices are Sensex and S&P CNX Nifty.

Financial market (Stock Market) is a platform for investors to own some shares of a company. Investors will then become a part of the company members and share in both gains and losses of that particular company. This is a better way for the investors to get extra income apart from their regular salary. Changes of share prices on every day make it more volatile and difficult to predict the future price. When purchasing a stock, it does not guarantees to give anything in return. Thus, it makes stocks risky in investment, but investors can also get high profit return. When investors take wrong decision in choosing the counters, it may end up in capital loss. The behavior of stock market returns has been deeply discussed over some years. In this paper, the hidden states and sequence are generated for stock market values using Hidden Markov Model (HMM) through software.

Review of related works

There are so many researches going on stock market analysis. Rabiner (1989) used precise HMMs, in which the state sequence estimation problem can be solved very efficiently by the Viterbi algorithm whose complexity is linear in the number of nodes, and quadratic in the number of states. However, this algorithm only emits a single optimal (most probable) state sequence, even in cases where there are multiple (equally probable) optimal solutions. Hassan and Baikunth Nath (2005) used HMM to predict next day closing price for some of the airlines. They considered four input attributes for a stock, and they were the opening price, highest price, lowest price and closing price. These four attributes of previous day were used to predict next day's closing price. Hassan (2009) introduced the new combination of HMM and Fuzzy model to forecast the stock market data. He classified the data set as daily opening, high, low and closing prices to predict the next day's closing price. HMM-fuzzy model is more reliable and profitable than the other model.

Jyoti Badge (2012) used Macro-Economic factor as a technical indicator, which is used to identify the patterns of the market at a particular time. For selecting technical indicator author was applying principal component

* Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India.
Email: sasikumarmsu@gmail.com and sheik.stat@gmail.com

analysis. Luca et al., (2013) investigated the dynamic patterns of stock markets by exploiting the potential of the HMM for defining different market regimes and providing transition probabilities for regime-switching. Tuyen (2013) used HMM to estimate the parameter of the Markov Black-Sholes model to predict the option prices in the stock market. The historical data of daily VN-Index (Vietnam Stock Market) were taken from 2009 to 2011 for finding the four hidden states corresponding to the Normal distribution $N(\mu_i, \sigma_i)$ for $i = 1,2,3,4$ with the help of HMM. Kai Cui (2014) explained that, the variation of financial time sequence for Shangai composite index was predicted by introduction of a dual state HMM. He also justified that the HMM was the best tool to predict the variation of financial time sequence. Somani et al., (2014) surveyed support vector machine, neural network and HMM in the area of stock market forecasting. HMM is more efficient in getting information from the result, showing future behavior of stock market valuesand fluctuations.

Methodology

Hidden Markov Model

HMM has been successful in analyzing and predicting phenomena's relying on a time dependence or time series. It is very effective and intuitive approach to many sequential pattern recognition tasks, such as speech recognition, protein sequence analysis, machine translation, pair wise and multiple sequence alignments, gene annotation, classification and similarity search.

A HMM is a doubly stochastic process in which an underlying stochastic process is unobservable, which means that the state is hidden. This can only be observed through another stochastic process that produces a sequence of observations. Thus, if $S = \{S_n, n=1, 2,...\}$ is a Markov process and $F = \{F_k, k=1, 2, ...\}$ is a function of S, then S is a hidden Markov process or HMM that is observed through F, and S is also known as the state process that is hidden and F as the observation process that can be observed. The observed event is called as a "symbol" and the invisible factor underlying the observation a "state".

A HMM is usually defined as a 5-tuple (S, F, P, ψ, π), where

$S = \{s_1, s_2, ...,s_n\}$ is a finite set of n states.

$F = \{o_1, o_2, ...,o_m\}$ is a finite set of m possible symbols.

$P = \{p_{ij}\}$ is the set of state-transition probabilities, where p_{ij} is the probability that the system goes from state s_i to state s_j.

$\psi = \{\psi_i(o_k)\}$ are the observation probabilities, where $\psi_i(o_k)$ is the probability that the symbol o_k is emitted when the system is in state s_i.

$\pi = \{\pi_i\}$ are the initial state probabilities; that is the probability that the system starts in state s_i.

As the states and the output sequence are understood, it is usually denoted by the parameters of a HMM by $\lambda = (P, \psi, \pi)$.

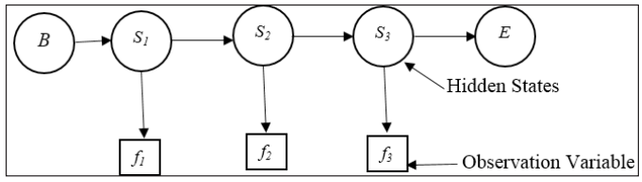


Figure 1: General Structure of a Hidden Markov Model

From the Figure 1, the $S_i$are the hidden states that is to be estimated and the F_i are the observation of the random variables from which the S_i are to be estimated. The letters B and E indicate the beginning and end of the sequence of states.

Transition Probability Matrix

The transition probability P_{jk},where $P_{jk} \geq 0$, for all j. These probabilities may be written in the matrix form,

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

This is called the transition probability matrix (tpm). P is a stochastic matrix ie.a square matrix with non-negative elements and row total is equal to one.

Materials and Methods

In this paper, Oil India Ltd is sample share, its daily close value data for two months period is considered. Three observing symbols I, N and D are indicated. The symbol I-indicates Increasing, N-indicates No change and D-indicates Decreasing. If n^{th} day's close value – $(n-1)^{th}$ day's close value > 0, then observing symbol is I. If n^{th} day's close value – $(n-1)^{th}$ day's close value < 0, then observing symbol is D. If n^{th} day's close value – $(n-1)^{th}$ day's close value = 0, then observing symbol is N.

Seven hidden states are assumed and are denoted by the following symbols S_1, S_2, S_3, S_4, S_5, S_6, S_7

where,

S_1 - very low

S_2 - low

S_3 - moderate low

S_4 - no change

S_5 - moderate high

S_6 – high

S_7 – very high

Since the above mentioned states are not directly observable, in this situation the stock market values are considered as hidden. From the hidden state sequences, it is possible to produce the observations.

The various probability values of tpm and emission probability matrix (epm) for difference in one day, two days and three days close values are calculated as follows: tpm and epm for one day close value difference

	S1	S2	S3	S4	S5	S6	S7
S1	0	0	0	1	0	0	0
S2	1/5	2/5	1/5	0	0	0	1/5
S3	0	0	0	1/11	8/11	2/11	0
S4	0	1/4	2/4	1/4	0	0	0
S5	0	1/14	7/14	1/14	3/14	2/14	0
S6	0	0	1/6	0	4/6	0	1/6
S7	0	1/2	0	0	0	1/2	0

Figure 1(a): tpm

Table 1: The Closing Value of a Stock Market Tpm and Epm for two Day Close Value Difference

S. No	*Close Value*	*Diff in 1 day*	*Observation Symbol*	*Diff in 2 day*	*Observation Symbol*	*Diff in 3 day*	*Observation Symbol*
1	458.00						
2	465.70	7.70	I				
3	469.00	3.30	I	-4.40	D		
4	467.35	-1.65	D	-4.95	D	-0.55	D
5	469.90	2.55	I	4.20	I	9.15	I
6	473.65	3.75	I	1.20	I	-3.00	D
7	466.00	-7.65	D	-11.40	D	-12.60	D
8	487.20	21.20	I	28.85	D	40.25	I
9	501.00	13.80	I	-7.40	D	-36.25	D
10	518.85	17.85	I	4.05	I	11.45	I
11	508.60	-10.25	D	-28.10	D	-32.15	D
12	500.15	-8.45	D	1.80	I	29.90	I
13	478.45	-21.70	D	-13.25	D	-15.05	D
14	478.45	0.00	N	21.70	I	34.95	I
15	467.50	-10.95	D	-10.95	D	-32.65	D
16	456.30	-11.20	D	-0.25	D	10.70	I
17	451.95	-4.35	D	6.85	I	7.10	I
18	457.05	5.10	I	9.45	I	2.60	I
19	459.25	2.20	I	-2.90	D	-12.35	D
20	458.50	-0.75	D	-2.95	D	-0.05	D
21	458.50	0.00	N	0.75	I	3.70	I
22	458.50	0.00	N	0.00	N	-0.75	D
23	453.65	-4.85	D	-4.85	D	-4.85	D
24	457.00	3.35	I	8.20	I	13.05	I

S. No	Close Value	Diff in 1 day	Observation Symbol	Diff in 2 day	Observation Symbol	Diff in 3 day	Observation Symbol
25	456.75	-0.25	D	-3.60	D	-11.80	D
26	467.65	10.90	I	11.15	I	14.75	I
27	470.10	2.45	I	-8.45	D	-19.60	D
28	470.95	0.85	I	-1.60	D	6.85	I
29	483.30	12.35	I	11.50	I	13.10	I
30	481.30	-2.00	D	-14.35	D	-25.85	D
31	483.45	2.15	I	4.15	I	18.50	I
32	478.65	-4.80	D	-6.95	D	-11.10	D
33	480.65	2.00	I	6.80	I	13.75	I
34	477.50	-3.15	D	-5.15	D	-11.95	D
35	485.95	8.45	I	11.60	I	16.75	I
36	487.55	1.60	I	-6.85	D	-18.45	D
37	486.00	-1.55	D	-3.15	D	3.70	I
38	492.60	6.60	I	8.15	I	11.30	I
39	491.15	-1.45	D	-8.05	D	-16.20	D
40	493.70	2.55	I	4.00	I	12.05	I
41	493.70	0.00	N	-2.55	D	-6.55	D
42	488.35	-5.35	D	-5.35	D	-2.80	D
43	490.00	1.65	I	7.00	I	12.35	I
44	499.00	9.00	I	7.35	I	0.35	I
45	501.25	2.25	I	-6.75	I	-14.10	D

	I	N	D
S1	0	1	0
S2	1/5	0	4/5
S3	10/11	1/11	0
S4	0	1/4	3/4
S5	5/14	1/14	8/14
S6	1/6	0	5/6
S7	0	0	1

Figure 1(b): epm

	I	D
S1	1	0
S2	1	0
S3	14/17	3/17
S4	0	1
S5	2/14	12/14
S6	0	1
S7	0	1

Figure 2(b): epm

	S1	S2	S3	S4	S5	S6	S7
S1	0	0	0	0	1	0	0
S2	0	0	1/4	0	1/4	0	2/4
S3	0	0	5/17	0	9/17	3/17	0
S4	0	0	1	0	0	0	0
S5	1/14	2/14	7/14	1/14	3/14	0	0
S6	0	1/3	2/3	0	0	0	0
S7	0	1/2	1/2	0	0	0	0

Figure 2(a): tpm

tpm and epm for three day close value difference

	S1	S2	S3	S4	S5	S6	S7
S1	0	0	0	0	2/4	1/4	1/4
S2	0	0	1/9	0	5/9	1/9	2/9
S3	0	0	3/7	0	4/7	0	0
S4	0	0	1	0	0	0	0
S5	2/16	6/16	3/16	0	5/16	0	0
S6	0	1	0	0	0	0	0
S7	2/3	1/3	0	0	0	0	0

Figure 3(a): tpm

	I	D
S1	1	0
S2	1	0
S3	5/7	2/7
S4	1	0
S5	5/16	11/16
S6	0	1
S7	0	1

Figure 3(b): epm

From the above TPM and EPM hidden states and sequence have been generated using MATLAB software. Difference of one day, two day and three day hidden states and sequence are given below respectively. From the sequence and states we can predict the future values of stock value.

1	Sequence:	D	I	D	I	I	D	D	I	D	I
	States:	S4	S2	S7	S6	S5	S5	S5	S3	S5	S3
2	Sequence:	D	I	D	I	D	I	D	I	I	I
	States:	S5	S3	S5	S3	S6	S3	S5	S2	S3	S3
3	Sequence:	I	I	I	D	D	I	D	I	I	I
	States:	S5	S5	S3	S5	S5	S2	S6	S2	S5	S3

Conclusion

Stock market values are unpredictable because of the variation of several factors. So there is no single method which can perfectly forecast the stock price values, HMM is no exception. Even though through this paper, the HMM model easily recognized three states of the stock market and also it was used to forecast the future values. In this paper, hidden states and sequences have been generated to identify, so that, we can easily identify the future states and also easily identify the sequence whether the next day value is increasing or decreasing and increasing/ decreasing level can also be observed. We can identify whether the increasing level is moderate, high or high or very high and also decreasing level whether moderate or low or very low. This is very useful for short term as well as long term investors.

Acknowledgement

The authors acknowledge University Grants Commission for providing the infrastructure facility under the scheme of SAP (DRS-I). The second author acknowledges UGC for financial support to carry out this research work under Basic Science Research Fellowship.

References

Hassan, M. R., & Nath, B. (2005). Stock Market Forecasting Using Hidden Markov Model: A New Approach, *Proceedings of the 5th International Conference on Intelligent Systems Design and Applications*, 192-196.

Hassan, M. R., Nath, B., & Kirley, M. (2007). A fusion model of HMM, ANN and GA for stock market forecasting. *Expert Systems with Applications,* 33(1), 171-180.

Hassan, M. R. (2009). A combination of hidden Markov model and fuzzy model for stock market forecasting. *Neurocomputing,* 72, 3439-3446.

Badge, J. (2012). Forecasting of Indian Stock Market by effective macro-economic factors and stochastic model. *Journal of Statistical and Econometric Methods*, 1(2), 39-51.

Cui, K. (2014). Analysis of Financial Time Sequence using Hidden Markov Models. *Journal of Multimedia,* 9(6), 810-815.

Angelis, L. D., & Paas, L. J. (2013). A dynamic analysis of stock markets using a hidden Markov model, *Journal of Applied Statistics,* 40(8), 1682-1700.

Ibe, O. C. (2009). *Markov processes for stochastic modeling.* Elsevier Academic Press

Rabiner, L. R., & Juang, B. H. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine January*, 4-16.

Rabiner, L. R. (1989). A tutorial on HMM and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257-286.

Somani, P., Talele, S., & Sawant. S. (2014). Stock market prediction using Hidden Markov model. *IEEE*, 89-92.

Tuyen, L. T. (2013). Markov financial model using hidden Markov model. *International Journal of Applied Mathematics and Statistics,* 40(10), 72-83.

Weigend, S., & Back, A. D. (1998). What drives stock returns? An independent component analysis. *In Proceedings of the IEEE/IAFE/INFORMS 1998 Conference on Computational Intelligence for Financial Engineering, IEEE*, New York., pp. 141- 156.

Zhang, D., & Xiaomin, Z. (2009). Study on forecasting the stock market trend based on stochastic analysis method. *International Journal of Business and Management,* 4(6), 163-170.

Article can be accessed online at http://www.publishingindia.com

Claim Analytics across Multiple Insurance Lines of Business

Ravi Chandra Vemuri*, Balaeswar Nookala*, Ramakrishnan Chandrasekaran*, Madhavi Kharkar, Sarita Rao****

Abstract

Claim loss payouts attribute most significantly to the overall costs of any insurance company and subsequently have a greater impact on its profits. Settling claims early, detecting fraudulent claims, increasing customer satisfaction and customer retention through recurring business have become increasingly important. The key to gain competitive edge is the ability to quickly and efficiently explore and understand data, settle claims and enhance customer experience. Claims processing, for any line of business in insurance (i.e. auto, life or health) is time consuming and labor-intensive involving multiple systems and several business units.

To address the above challenges and support the claim handler's decision, information that is available during the first notification of loss (FNOL) is used to predict claim severity using predictive analytics. A claim that is straight forward with reliable evidences can be considered as a simple claim which can be settled quickly and dealt by a junior adjustor. While a claim which involves accident, death or a police case can be considered as complex and needs a team of senior adjustors or lawyers to get involved. Using the prediction model, the claims department can classify the claims based on its severity and assign it to the respective team thus improving its operations.

Additionally the research also involves analysis of similarities and differences between the key attributes across three lines of business in insurance (auto, life and health) that impact the claims severity. By further studying the claim trends across these lines of business for a particular geography or demography, business can further refine the risks considered during underwriting or can design new products with add-ons which will be beneficial to the customers and contribute to increased business.

Keywords: Insurance, Claim Analytics, Claim Severity, Fraud Detection, FNOL, Customer Satisfaction

Introduction

Claim analytics lets insurance companies view individual claims in context with total volume of claims received. Along with business rules, the power of combining numerous variables with statistical tools and processes can enable insurance companies to measure the severity of a claim and predict fraudulent claims early. Based on the results, claims with high severity can be handled by experienced personnel with improved turnaround time leading to effective claims management.

Information available during claim evaluation is different depending on the line of business. However, they can be classified against four profiles as explained in Table 1.

Table 1. Profiling Factors

Profiling	*Line of Business*		
	Life	*Health*	*Auto*
Custom-er	Old or new customer	Age of the Insured	Driver experi-ence and age
Claim	Claim Amount	Loss date and reporting lag	Loss date and reporting lag
Loss Details	Cause of death	Medical details	Cause and type of loss
Risk	Product type	Coverage amount	Vehicle capac-ity and make

Some of the parameters listed in the above table are considered across lines of business (LOB). During the

* CGI, Bangalore, Karnataka, India. Email: ravichandra.vemuri@cgi.com, E-mail: balaeswar.nookala@cgi.com, E-mail: ramakrishnan.chandrasekaran@cgi.com
** CGI, Mumbai, Maharashtra, India, E-mail: madhavi.patil@cgi.com, E-mail: sarita.rao@cgi.com

research, additional parameters from the source data in each LOB were used as required.

Using predictive analytics on the claims, relationship between the available information (i.e. input attributes) and the claim severity (i.e. depicted in the form of a score) is arrived so that it can be applied to future claims and most probable severity of the new claim will be known earlier in the process thus helping in effective claim management.

Data (Source, Quality & Volume)

Claims data for each line of business (i.e. life, health and auto) have been used to build the respective Claims Severity Scoring Model. Internal data which is usually collected by the insurance company when a policy is issued and while a claim is accepted from the customer has been taken into consideration to build the model.

With experience and knowledge on the insurance business and with statistical analysis on available data, the business analysts and the technical team built a good dataset of more than 75000 claim records for life, 34000 records for auto and 13000 records for health insurance.

Incomplete, incorrect, junk and duplicate records were removed from the claim records. Wherever relevant, inconsistent or missing data were rectified using business rules.

Dataset was further divided into 3 parts based on claim year:

- 1st set (training dataset) - Used to build 1st version of the model
- 2nd set (validation dataset) - Used to validate the model and then recalibrate
- 3rd set (testing dataset) – To implement the model and arrive at "Claim Severity Score" as well as different complexity levels.

Data Boundaries were as follows:

- Additional data volume would help in multiple cycles of recalibration and build confidence on model output.
- Volume of data available for a few categories of vehicles make or products was high while it was very low for other categories
- Number of levels available for few of the categorical variables like place of accident was very high and patterns could not be studied further
- Data for particular age group or a sector was completely missing in the dataset
- All claims are within the policy effective date

Target Field And Input Fields

Target Field

Output of the prediction model is the claim severity which will be depicted in the form of a score for each claim. Higher the score more severe or complex is the claim. This information was not present in the input data set. So for the training data set, the target 'Y' is arrived based on few business assumptions as listed in Table 2:

Input Fields

All the relevant fields from the claims and policy data that would have an impact on the target or dependent variable "Claim Severity Score" ("Y") are taken into account. Few input or independent variables are derived from the existing fields which could be used further to build the model.

Table 2: Assumptions

LOB	*Auto Insurance*	*Life Insurance*	*Health Insurance*
Assumptions for high scoring	Vehicles carrying heavy goods	Product type covering all family members	Higher claim amount than the average claim amount
	Vehicles moving across states and on hilly roads	Reporting delay	Reporting delay
	Vehicles driven by other than owner	Higher claim amount requested for a certain product type	Higher doctor consultation charges
	Passenger death due to accident	Higher age at issue	Higher age at issue

For e.g. in auto insurance, "Claim Reporting Delay" was derived using "Claim Reporting Date" and "Accident Date". "Claim Reporting Delay" was included in the dataset while "Claim Reporting Date" and "Accident Date" fields were excluded.

Both categorical and continuous attributes were considered for building the model. Table 3 below lists out some of the independent variables that were considered for each line of business

Methodology

Tools

- Microsoft SQL Server for data cleansing
- R-Rattle to build, validate, recalibrate and implement the model
- Minitab for statistical analysis

Predictive Modeling

The training data set was used to build the prediction model. As the target "Y", i.e. the "Claim Severity Score", is a continuous numeric variable, Linear Regression was the most suitable model to be implemented since it considered maximum number of the input variables to build the relation with "Claim Severity Score". Output of the model is depicted in the Table 4.

R Square Value for auto and life insurance threw a good confidence level (>75%). As the volume of health data was low and the distribution of data across different products was not significant the confidence came in low.

Table 3: Independent Variables

Variables / LOB	*Auto Insurance*	*Life Insurance*	*Health Insurance*
Independent Variables (X1,X2,X3, X4….)	• Sum Insured • Vehicle Class • Vehicle Capacity • Driver Age • Driver Experience • Claim History • Loss Reason	• Region • Underwriting Category • Policy Face Amount • Occupation at Issue • Product type	• Gender • Product type • Total Amount Claimed • Sum Insured • Doctor Consulting Charges • Network Hospital
Dependent Variable (Y)	Claim Severity Score		

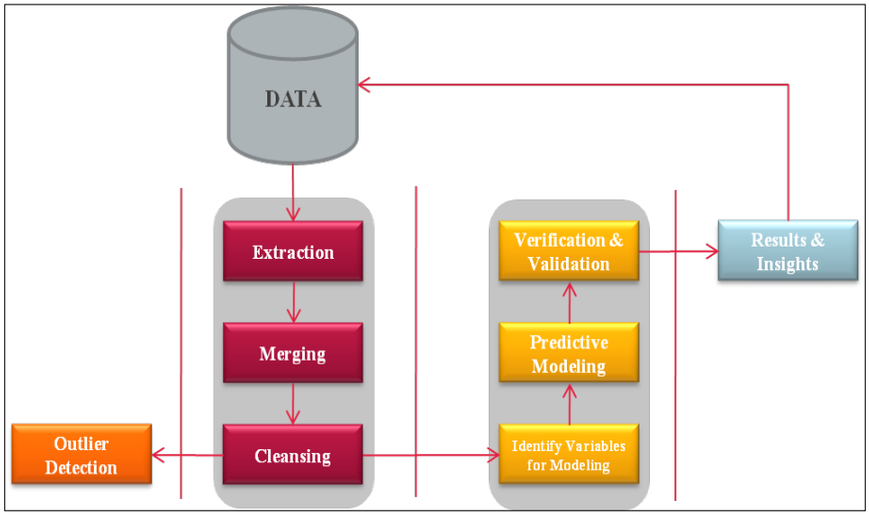


Figure 1: Data Analytics Flow

Table 4: Linear Regression Output

LOB	*Auto Insurance*	*Life Insurance*	*Health Insurance*
Linear Regression (Numeric)	R–Square Value of 0.84**	R-Square value of 0.77	R-Square value of 0.48

Additionally, the variance analysis output (ANOVA) given by the linear regression model was analyzed for each independent variable. The independent variables that did not show a strong relationship (p value >0.05) were excluded and the model was rebuilt.

For e.g. in auto insurance, we observed that one of the fields with high 'p' value was 'Summons Type'. From business point of view, this is an important attribute to be considered, as litigation costs would have increased the complexity of the claim. But as the historic data does not have enough claims of this type or the data was not captured accurately, the model suggests a weak relationship with the severity score.

Model Validation & Recalibration:

Model built in the above step was applied on the next set i.e. validation dataset. It was observed that except for few outliers, the observed and predicted value was almost the same.

As the validation showed a significant relationship between observed and predicted value, prediction model was further recalibrated by adding the validation data also to the model. Outliers were removed before building the model. This step helped to further strengthen the model with additional claims which were of similar nature or had different patterns which were not available in the first dataset.

Additionally variation of few of the input variables was analyzed to validate the business relationship.

The above chart (Figure 3.a) shows that the severity scores are higher if the 'Driver's Experience' is less than 1 year or is greater than 6 years. However, as the score has been derived using multiple input parameters the variation may

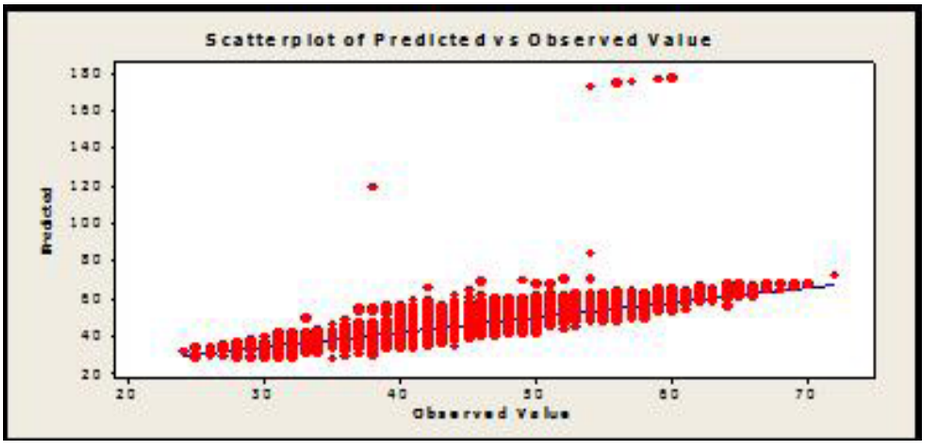


Figure 2. Scatter plot of Predicted vs. Observed value

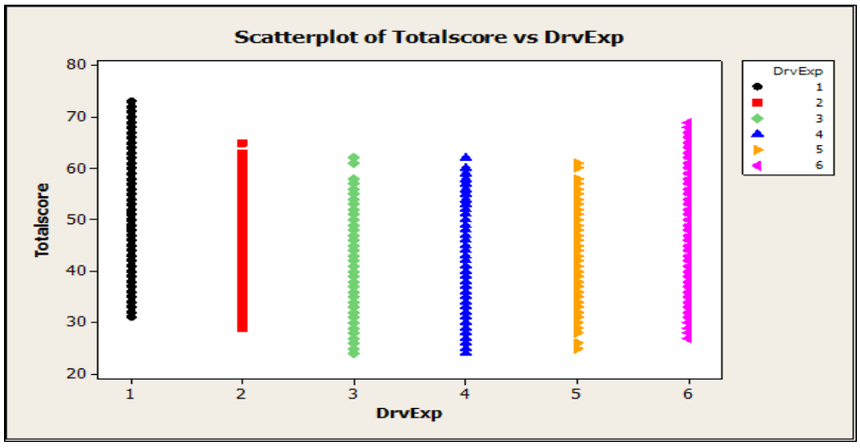


Figure 3.a. Total Score vs. Driver Experience

also depend on other factors. Hence, second parameter was picked up for analysis as shown in Figure 3.b.

The 'Nature of Loss' shows high variation and higher score in case of external accident or vehicle theft. With above analysis we have a higher confidence on the behavior of the prediction model as it is in line with the theory. Similar analysis was done on life data to validate output of the model.

Model Implementation

Linear regression model was then applied on 3rd dataset (testing data) to arrive at the Claim Severity Score. Scores were further classified into different levels of complexity using statistical techniques like box-plot (Figure 4) that helps us understand the distribution of the data.

Table 5 provides classification arrived for each line of business.

ANALYSIS RESULT

Based on the above distribution, claims management can assign the set of claims to respective level of adjusters. Claims marked as complex or very complex can be further analyzed for fraud. Additionally claims with a very low score can be automated after consulting with similar line of business and settled immediately.

As additional data becomes available at later stages in the claim lifecycle and are also validated by business, the model should be recalibrated using relevant information to derive an updated severity score and complexity.

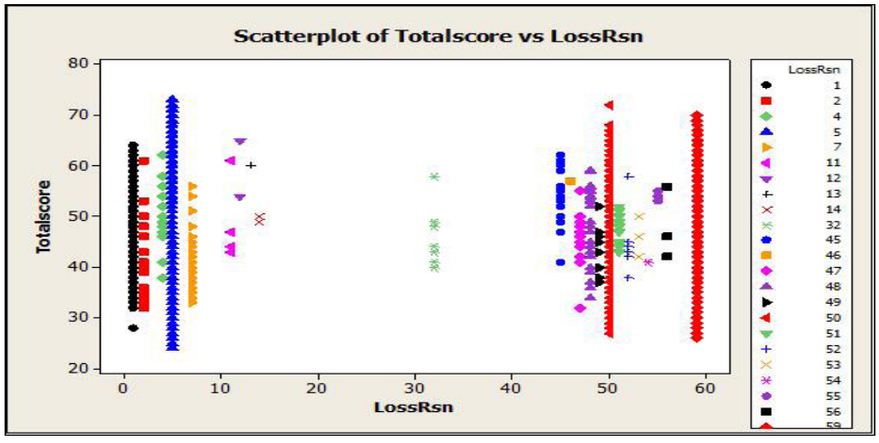


Figure 3.b. Total Score vs. NOL

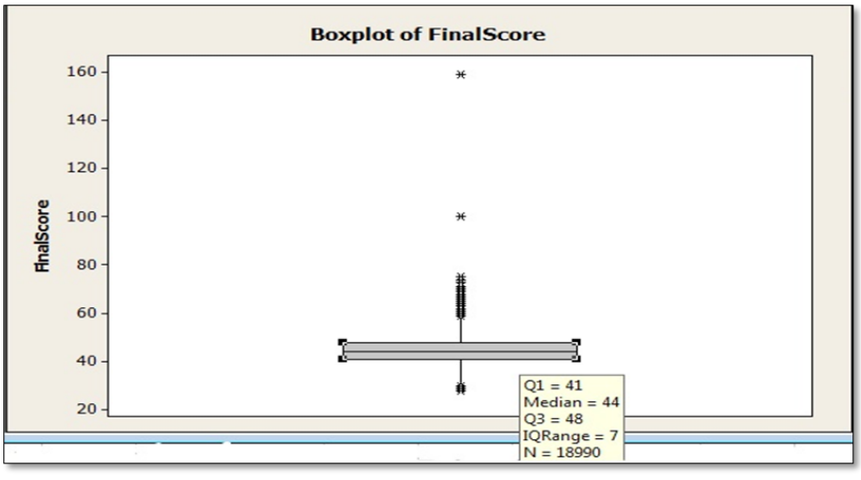


Figure 4. Box-plot of Final Score

Table 5. Distribution of Claim Complexity based on Severity Scores

Complexity Levels	*Auto Insurance*		*Life Insurance*		*Health Insurance*	
	Claim Severity Score	*% of data*	*Claim Severity Score*	*% of data*	*Claim Severity Score*	*% of data*
Simple	<41	20%	<14	13%	<14	18%
Medium	41-44	34%	14-17	53%	14-16	60%
Complex	45-48	24%	18-23	32%	17-19	19%
Very Complex	>48	22%	24-28	2%	>20	2%

Table 6. Key Attributes

	Auto Insurance	*Life Insurance*	*Health Insurance*
Renewal Frequency	Usually 1 year	More than 1 year	Usually 1 year
Approximate claims per month (India)	10 to 12 thousand	70 to 80 thousand	4 to 5 thousand
Number of independent variables having possible impact on claim severity	15	6	10
Attributes showing strong relation with Claim Severity Score	Driver Experience Driver Age Nature of Loss Claim History Reporting delay Vehicle Capacity and Make	Product Type Cause of Death Claim Amount Sum Assured Underwriting Category	Age of Insured Claim Amount Surgery and Consulting Charges Investigation Network Hospital Bonus Sum Insured
Attributes not showing relation with Claim Severity Score	Summons Type NCB Antitheft Endorsement	Premium amount paid Annual Income Reporting delay	Policy years Co-payment applicable Pre and Post hospitalization expenses Other non-hospital expenses Miscellaneous charges

Further comparing the observations from each line of business, Table 6 below summarizes some of the key attributes that reflect the significance of the claims severity score for these businesses.

From the above table, we observe that age factor is one of the parameters that has appeared to be common between both auto and health insurance. Additionally the claim amount is also an important input factor that determines severity in life and health insurance. Hence, we notice few similar trends across lines of business.

However, due to the differences in the nature of each of the lines of business, time at which insurance can be claimed etc. we find that there are many factors that are dissimilar across lines of business.

By further filtering the claims data for a particular geography or demography, claims severity scoring can be done across the lines of business. Based on the trends observed in high complexity claims across customer profile or claim profile, underwriters can get more inputs in defining additional risks. Also, by studying the reason for high volume deaths (due to a disease) in the particular geography through life insurance claims, health insurance products can be re-designed and premium discounts can be provided for people who are taking precautionary measures. This in turn will help increase business of health insurance products, reduce life insurance claims and promote well-being of the society.

CONCLUSION

A seamless claims management system is strengthened by using the claims data effectively. Higher the volume of data, greater the opportunity to see through, analyze, track patterns, understand claim(s) behavior, etc. and greater the potential in detecting fraud and staged claims.

In essence, claims analytics can provide insights around the below business imperatives

- Claim Segmentation and Service Prioritization
- Return on Investment
- Customer Relationship
- Product Innovation
- Continuous Improvement Process

Learnings and Next Steps

- Severity score and complexity predicted by the model to be validated by business users to increase the confidence on prediction.
- Challenges faced during data cleansing process to be further discussed with business to improve the quality of data.
- Data to be further stratified to analyze patterns and behaviors of important categories like the particular vehicle class code- e.g. commercial or trade vehicles for auto insurance.
- Based on the volume of the data, model scores can be re-calibrated on yearly basis or as and when required (i.e. when new patterns are seen in industry or there is deviation in the model to arrive at the range of the complexity levels to determine any changes to the complexity bands suggested above, etc.)
- Model to be further enhanced by considering external data like credit score of individuals, information from social networking sites, etc.
- From data collection point of view, level of data captured for the attributes that are showing strong relationships can be further streamlined e.g. Loss Reason Code should be captured under the defined categories rather than listing it as 'Miscellaneous' or 'Others'. This will help in arriving at more appropriate scoring for the claim.
- We have arrived at the severity score and found the most important parameters across the 3 lines of business. This could be further used to build fraud detection model.

REFERENCES

Simonson, E., & Jain, A. (2014). *Analytics in Insurance.* Everest Group Research, Genpact.

Lentz, W. (2013). *Predictive Modeling – An Overview of Analytics in Claims Management*. GenRe Research.

Banerjee, D., & Dasgupta, A. (n.d.). *Claims Predictive Modeling*. Deloitte.

Bari, A., Chaouchi, M., & Jung, T. (2014). *Predictive Analytics for Dummies.* A Wiley Brand.

Article can be accessed online at http://www.publishingindia.com

Delay Prediction of Aircrafts Based on Health Monitoring Data

B. A. Dattaram*, N. Madhusudanan,*

Abstract

Flight delay is a major issue faced by airline companies. Delay in the aircraft take off can lead to penalty and extra payment to airport authorities leading to revenue loss. The causes for delays can be weather, traffic queues or component issues. In this paper, we focus on the problem of delays due to component issues in the aircraft. In particular, this paper explores the analysis of aircraft delays based on health monitoring data from the aircraft. This paper analyzes and establishes the relationship between health monitoring data and the delay of the aircrafts using exploratory analytics, stochastic approaches and machine learning techniques.

Keywords: Aircraft Delay, Faults and Alerts, Markov Chains, Time before Failure, Stochastic Ensemble

Introduction

The failure of an aircraft to take off due to component faults and unplanned maintenance leads to revenue loss, affecting the core business of the aircraft owning companies. The problem with component faults is significantly observed in the ageing aircrafts. It is, therefore, necessary to anticipate delays so that proper maintenance processes can be initiated before an actual delay occurs. The health of the aircraft is monitored through fault and alert messages which are relayed from the different subsystems, during its journey. These faults and alerts are leading indicators of the health of the aircraft. We, in this paper explore the relationship between aircraft delays and the fault and alert messages from the aircraft.

Aircraft Sub-systems Description and ATA Codes

An aircraft can be considered as an assembly of different sub-systems like Engine, Hydraulics, Cockpit, Landing Gear, Electrical Components etc. which work cohesively. The industry standard body has established codes for each of the sub-systems. Table-1 describes Air Transport Association of America (ATA) chapter numbers for some of the important sub-systems of the aircraft.

In this paper terminologies like Sub-Systems, Airframe Systems and ATA Chapters are interchangeably used. The complete list of ATA Chapters and corresponding codes is provided in.

Table 1: List of Airframe Systems and Corresponding ATA Numbers

ATA Number	*ATA Chapter Name*
ATA 23	Communications
ATA 22	Auto Flight
ATA 24	Electrical Power
ATA 28	Fuel
ATA 72	Engine – Reciprocating

Data Description

In this section, we will describe the data that we have used for Exploratory Analysis and Time-to-Failure Prediction Model. The data consists of two separate sets: the first set contains the alerts and faults obtained from all aircrafts during their flight journey and the second set contains the delay dates of all aircrafts.

Faults and Alerts data consists of the Aircraft Number, ATA Chapter code associated with the fault or alert message and also the Timestamp. Table-2 provides a snapshot of Alerts

* IBM India Pvt Ltd, Bangalore, Karnataka, India. E-mail:dattarao@in.ibm.com, mnaraya7@in.ibm.com

and Faults. These are the health monitoring messages relayed by different aircrafts during the journey and may not always indicate an imminent failure.

The Delay Table consists of delay or cancellation events along with the Aircraft Number, Timestamp and associated ATA Chapter code. Delay events represent situations when an aircraft failed to take off. In this paper, we establish the relationship between the Faults and Alerts data and delays or cancellations (hereafter also referred to as Failures) using exploratory and machine learning techniques. In this paper, alerts and faults have been treated homogenously. Delays and cancellations have been treated together as failure events. Data for a total of 63 aircrafts all belonging to the same series with same engine was considered. The age of the aircraft is in the range of 15-20 years. A total of one year of data was provided.The total number of delay events were approximately 3.5% of the total take-off events.

Table 2: Format of Alerts and Faults Data Obtained from Different Aircrafts

Aircraft Number	*ATA Chapter Code*	*Message Type*	*Timestamp*
A1	32	Alert	2012/04/05 14:11:05
A2	73	Fault	2012/04/05 14:30:25
A23	73	Alert	2012/04/05 13:11:05
Event Type	Aircraft Number		Timestamp
Delay	A1		2012/04/05
Cancellation	A1		2012/04/12
Delay	A2		2012/05/05

Related Work

In the field of aerospace, a lot of work depends on reliability theory and preventive maintenance. In particular, there is lot of effort towards data mining based prediction for component replacements (Létourneau, Fazel, & Matwin, S. 1999). These approaches also make use of the sensor data obtained from the aircraft along with the text comments from maintenance crew to predict the most optimal time to replace a component. Data mining methods like Naïve Bayes, decision trees or even hybrid models have been used earlier to predict the most optimal time for component replacement.

Delay Prediction: Our Approach

In this section, we will describe our approach on data analysis and delay prediction using machine learning techniques. In the next few sections we describe our insights and results from exploratory analysis, Hidden Markov Models and prediction of Time Before Failure (TBF) using a combination of regression trees and stochastic ensemble modeling.

Exploratory Analysis

The objective of exploratory analysis is to establish that there is information contained within the ATA Fault and Alert messages for the prediction of the next failure event and the consolidation of additional information which can prove useful for understanding underlying patterns in the occurrence of messages or for modeling TBF. In this method, we have analyzed the delay events and the

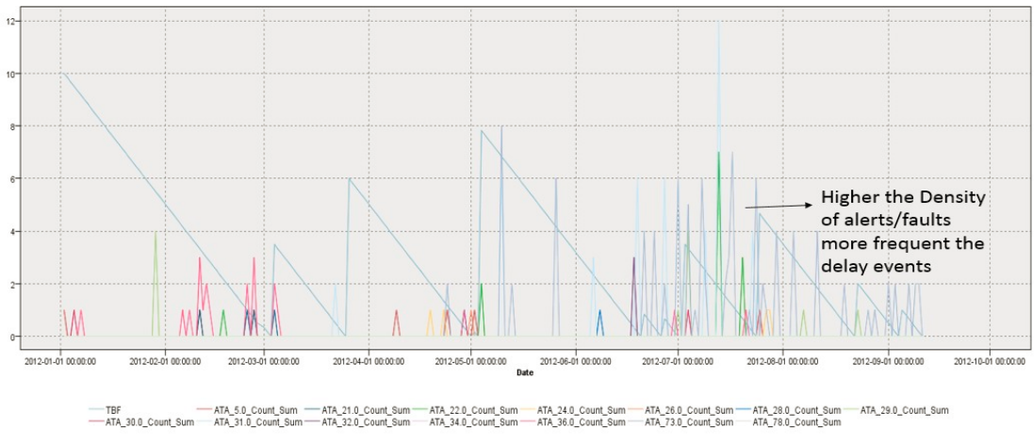


Figure 1: Plot of ATA Count vs. Delay Events for One of the Aircraft

ATA code occurrences based on the count of each ATA aggregated on daily basis.

Figure 1 plots the aggregated count of each ATA message per day against the delay events and TBF. This plot allows us to make observations which reveal the relationship between alert and fault messages of different ATA codes and the failure events.

It can be observed that the density of ATA alerts or faults is closely related to the delay events represented by the saw-tooth waveform. A higher amplitude of the saw-tooth represents a longer time before another failure event occurred. At the beginning of the year, the alert and fault message density is low and so are the delay events. When the density of alerts and faults increases, the TBF value also drops leading to a delay event. This visualization allows us to establish a causal relationship between alerts and faults messages and delay events due to different sub-system issues.

Based on insights revealed by the above trend chart, further analyses of daily message count and their relationship to TBF was performed for each ATA message, across all aircrafts. Figure 2 shows the results for one of the ATA codes.

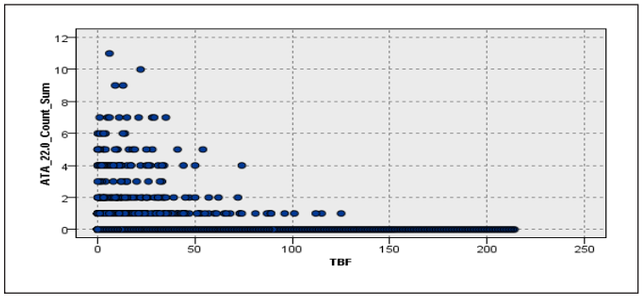


Figure 2: Per day Count of ATA 22 Messages vs. TBF

The interpretation of the above graph is as follows - as the daily count of ATA 22 messages increases, the distribution of the TBF shifts closer to zero. In other words, a higher frequency of ATA 22 messages indicates a higher temporal proximity to the next failure event. It was found that the pattern is similar across all faults and alert messages of each ATA - a higher frequency indicates a closer proximity to failure.

Hidden Markov Model

HMMs are two layered stochastic models with one hidden layer which cannot be observed. The hidden layer Markov process can be observed through the observations or symbols which hidden layers emit in agreement to the laws of probability and Markov chains. HMMs are widely used in gene sequencing and speech recognition.

Why HMMs for Aircraft Delay Prediction?

In this paper, we have made an attempt to model the aircraft as a two layered Markov chain. The hidden layer represents the interaction of different aircraft sub-systems and is modelled as a first order Markov chain. This interaction among different sub-systems is complex to specify and model and hence is considered as a hidden layer. The top layer which can be observed in the form of alerts and faults messages with specific ATA codes is also treated as a first order Markov chain. It can be seen that the interaction of the sub-systems in the hidden layer emits the observations or the messages with ATA codes. This is depicted as a schematic diagram in Figure 3.

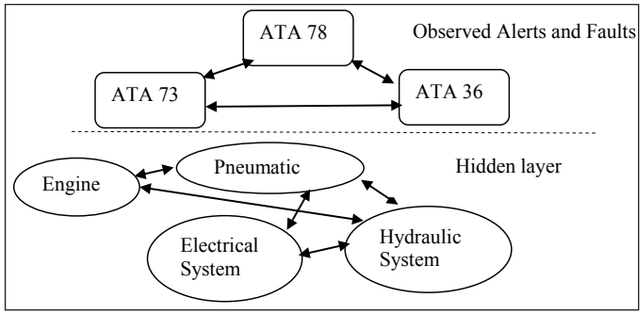


Figure 3: Schematic Diagram Showing Hidden and Observed Layers

HMM is proposed as a model with the following objectives:

1. HMM Model on aircraft delay data should help in understanding the interactions among the different sub-systems. Like, how tightly coupled are certain sub-systems or are there sub-systems which do not influence others
2. HMM based Delay Prediction Model

HMM Specification

A discrete HMM model is specified using emission and transition probabilities. Emission probabilities is defined as the probability of an observation or a symbol being emitted from a hidden state. Transition probabilities include the probability of transition from one hidden state to next hidden state.

In order to achieve the above objectives, we define the HMM with 4 hidden states and 16 symbols which compromise all the ATA codes and the delay event represented as symbol "F" (Salfner, 2005).

Observations={"5","21","22","24","26","28","29","30","31","32","34","36","45","73","78","D"}

Hidden States = {M1, M2, M3, DELAY}

Training the HMM

Once the model is initialized with emission and transition probabilities, the model is trained using Baum-Welch algorithm to derive the most likely set of transition and emission probabilities. 80% of the data is used as input to Baum-Welch algorithm to derive the emission and transition probabilities.It is ensured during initialization that only the Failure State emits the Delay or "D" symbol. By initializing only the Failure State to emit the Delay symbol, it can be guaranteed that failure symbol in training sequence will change the transition to Failure State (Salfner, 2005). For details of theory and background on Viterbi and Baum-Welch algorithms, refer to (Salfner, 2005; Rabiner & Juang, 1986)

Results from HMM model

As mentioned earlier, one of the objectives of the HMM model is to understand the interactions among the different sub-systems of an aircraft. In this regard, the Table 4 lists the emission probability from each of the hidden states and Table 5 lists the transition probabilities.

From the emission probabilities, it can be observed that hidden state M1 has maximum emission probability only for Indication and Recording sub-systems while M2 has the most number of Protection sub-systems along with Landing, Pneumatic and Navigation. The maximum emission probability for each of the ATA codes and the hidden state reveals how some sub-systems in the aircraft are more tightly coupled. It also explains the sequence of ATA codes that are observed in the alerts or fault messages.

Table 4: Hidden State and Emission Probabilities

Hidden State	*Observations with maximum Emission Probability for the Hidden states*
M2	Maintenance Checks(5) , AC and Pressure (21), Electrical Power(24), Fuel(28), Auto Flight (22), Rain protection(30), Fire protection(26), Landing(32), Pneumatic(36), Navigation(34), Diagnostic(45), Exhaust(78), Hydraulic(29)
M3	Engine (73)
M1	Indication or Recording (31)
DELAY	Delay (D)

Table 5: Transition Probabilities Filled up to 3 Digits

	M1	*M2*	*FAILURE*	*M3*
M1	0.77	0.15	0.037	0.036
M2	0.015	0.85	0.103	0.027
FAILURE	0.035	0.88	0	0.079
M3	0.009	0.102	0.038	0.84

The transition probabilities reveal that M2 hidden state has transition probability of 0.103 to Failure state while other states have very low probability of transition to Failure state.

Predicting the Probability of the Failure State

Given an observation at time t, the next hidden state at time t+1 can be estimated by multiplying the posterior probabilities at time t and the transition probabilities obtained from the HMM. Posterior probability represents the marginal distribution of a state given by the observations. Results from running the test data of different sequence of ATA codes reveal that HMM predicts the Failure state with high accuracy (80%) when the underlying hidden state sequence is dominated by state "M2". But HMM fails to predict the Failure state when the hidden state sequence is dominated by "M3" i.e. the observed ATA codes are from Engine sub-system.

Failure Prediction Model Using Stochastic Ensembles

This section outlines the methodology underlying the stochastic ensemble model for predicting the Time Before Failure (TBF). This method is predicated on an approach where each ATA message occurrence / event is regarded to contain information which allows us to estimate TBF with some level of certainty. However, a combination of such estimations from multiple events can also be used to augment the final estimate. In other words, the frequency of each ATA message and equivalent metrics can be used as a stochastic predictor of TBF and these probability distributions can be progressively combined to arrive at a more deterministic estimation of TBF.

The modelling methodology adopts three stages –

1. Regression Tree Modelling – Uses clustering to establish time boundaries and derive frequency metrics derived for each ATA message event and build multiple regression tree models to predict TBF
2. Distribution Fitment – Fits an appropriate distribution to describe the spread of TBF in each leaf of the regression tree
3. Distribution Readjustment and Stochastic Ensemble – Readjusts the fitted distributions for earlier fault/alert message events to account for the time elapsed since those events occurred. Combines all readjusted TBF distributions observed since the last failure event to obtain an accurate estimate of TBF

The following subsections outline each of these stages in greater detail, along with results obtained for the same. This paper restricts itself to a description of the core methodology and does not detail out programming intricacies such as techniques used for handling NULL or NA values generated during data processing.

Regression Tree Modeling

Regression trees are decision trees built for the prediction of continuous variables. The basis of establishing a statistically significant variation in the distribution of TBFs across node splits is determined using the ANOVA method.

Based on the insights from exploratory analysis, the predictors used for TBF prediction have to be related to the frequency of messages which is calculated as F = N/T, where N is the number of messages over a time span T. Daily message count aggregates used for exploratory analysis are not ideal because daily boundaries were defined for convenience and do not reflect inherent time boundaries present in the data. Therefore, new time boundaries were derived from the data using a simple heuristics based univariate clustering method.

Messages for each ATA for each Aircraft were processed in chronological sequence. A new data point is retained in the same cluster if it falls within 1.96σ deviations from the mean of the cluster. If the aforementioned criterion fails, the data point is classified into a new cluster. Sample results for ATA message73 for a specific aircraft are shown in Figure 4. The diagram represents each occurrence of a message with a dot, and the vertical lines represent the beginning of a new cluster.

For any message event, let $(N_{ATAj})_{Ti\text{-}Ti}{}^{c}$ represent the number of messages of ATA code ATAj occurring between Ti^{c} - the time which represents the beginning of the current cluster to which the event belongs, and Ti - the time of the message event. Then the frequency computed for the event is given by –

The frequency of messages is calculated for two scenarios - (a) by treating each ATA message as a separate event stream (b) by treating all messages as a single message

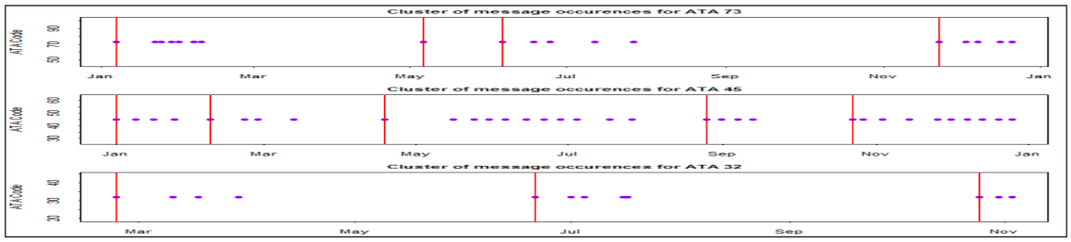


Figure 4: A Sample of Cluster Boundaries which were Extracted - ATA message 73 for one specific Aircraft

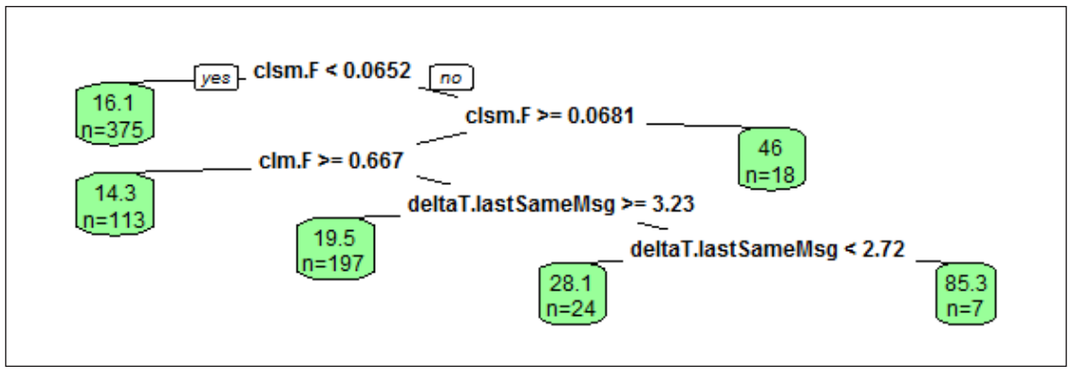


Figure 5: Regression Tree Built for ATA 5 Message Events

stream irrespective of the ATA code. This yields two different frequency metrics.

The regression tree models are built for each ATA code, across aircrafts, resulting in 15 regression tree models. One such regression tree built for ATA 5 is shown in Figure 5. The following predictors were computed for all aircrafts and considered for the regression tree model –

(i) Cluster frequency computed for each ATA message stream for every event (clsm.F)

(ii) Cluster frequency computed for all message events disregarding ATA codes (clm.F)

(iii) Time elapsed since the last message of the same ATA code (deltaT.lastSameMsg)

(iv) Time elapsed since the last message of any ATA code (deltaT.lastMsg)

Distribution Fitment

Each leaf of the regression tree model in Figure 5 displays the average Time Before Failure (TBF) of values in that leaf. However, each leaf (hereafter referred to as Class) contains a spread of TBF values which represent a probability distribution for TBF. This section discusses the fitment of an appropriate distribution to model the TBF for each leaf (Class).

From an examination of Cullen and Frey Graphs plotted for the TBF distributions, we can conclude that a gamma function is an appropriate choice for modelling the TBF distribution across all ATAs and Classes. Figure 6 shows a sample set of Cullen Frey Graphs plotted for ATA 45 for Classes 1 and 2. The graphs show that most of the observations fall close to the dashed line characterized by a gamma distribution.

Figure 7 shows the results of a sample fitment for ATA 73 Class 2. The results of each fitment yield two parameters for the fitted gamma distribution – shape and rate.

The final result of distribution fitment is a lookup table of values for each ATA code and Regression Tree Class

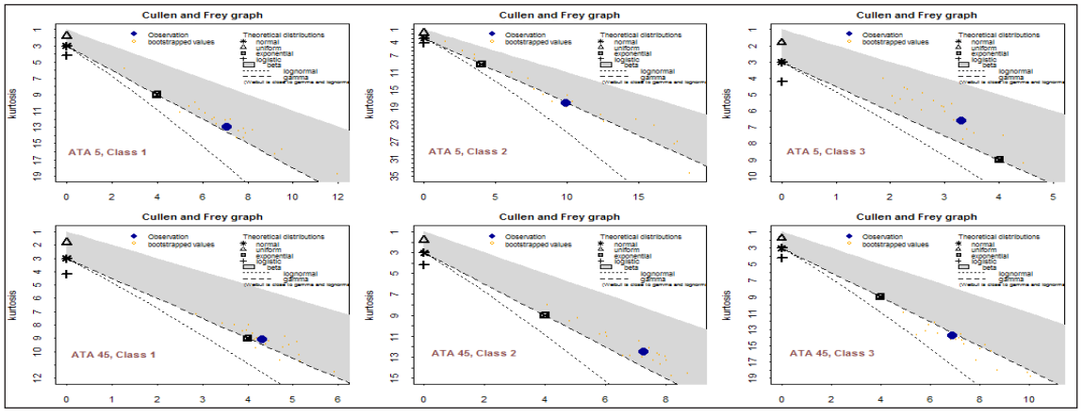


Figure 6: Cullen Frey Graphs for ATA 45 Classes 1 & 2

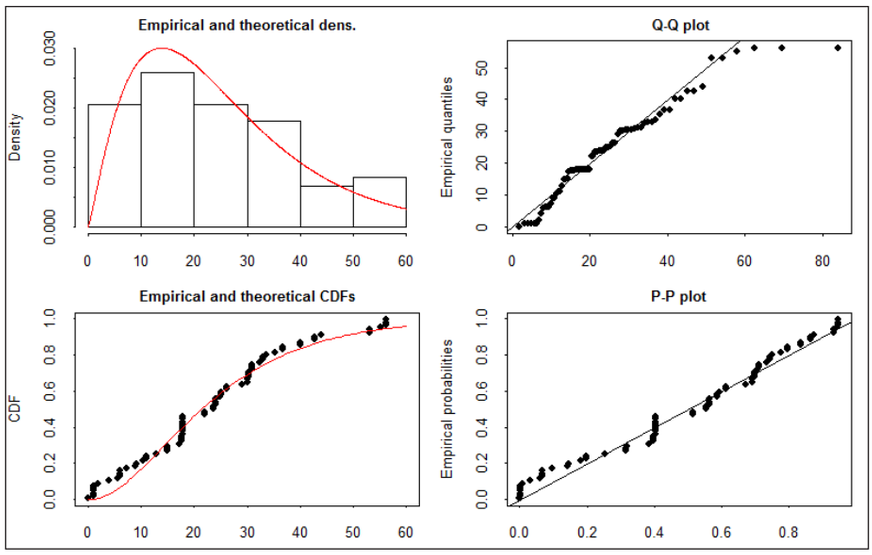


Figure 7: Gamma Distribution fitted for ATA 73 Class 2

for describing the distribution of TBF. A snapshot of this table is shown in Table 6.

Table 6: Snapshot of Distribution Fitment Lookup Table

ATA Code	*Class*	*Shape*	*Rate*
73	1	1.161687	0.08051
73	2	2.24427	0.089735
73	3	0.993364	0.065011
73	4	0.887306	0.057574
45	1	1.038601	0.071375
45	2	0.76693	0.029999
45	3	0.973843	0.055013
36	1	0.904286	0.04644
36	2	1.005415	0.061267
36	3	0.724237	0.043222
…	…	…	…

Distribution Adjustment and Stochastic Ensemble

This section describes the technique behind aggregating the stochastic predictions made by each ATA message event to provide an integrated prediction of Time Before Failure (TBF).

Integral to the technique of Stochastic Ensemble of predictions from occurrence of events is the readjustment of probability distributions. In this application, since the prediction made is the time to failure, the elapse of time allows us to invalidate a subset of the TBF distribution obtained from an event in the past. For example, if 6 days have elapsed since the last ATA 73 message which fell under Class 2, we know that the p(TBF < 6) = 0.This information can be used to progressively readjust the distribution through the passage of time.

Figure 8 illustrates the process of distribution readjustment of the probability distribution function and the cumulative distribution function after an elapsed time period of 6 days post the occurrence of the original event.

Stochastic ensembles are generated using the following process steps.

(i) The frequency estimators for each ATA event are used to predict the Class based on the Regression Tree Model

(ii) TheTBF distributions for each ATA and Class since the last failure event are obtained from the lookup table and adjusted for time elapsed

(iii) The readjusted distributions are combined using the stochastic ensemble method

The methodology adopted for combining the readjusted distributions has to satisfy the preconditions of being able to mimic the behavior of a deteriorating system and introducing greater certainty in the prediction of TBF. In addition, the information contained in distributions which predict a lower TBF need to be retained in the ensemble process. These properties are satisfied by considering a

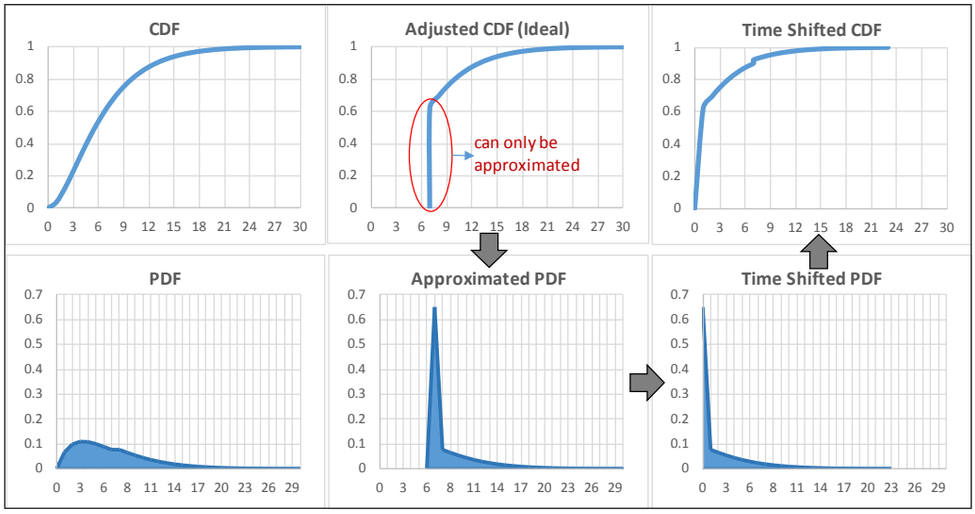


Figure 8: Distribution Readjustment Example (Left) Original Distribution at T=0 (Middle) Redistribution After T+6 days (Right) Final Distribution Time Shifted for 6 Days

product of the inverse cumulative distribution functions. If x_i is a random variable that denotes the uncertainty in TBFs predicted by past ATA message events, where $i \in \{1,2,3,\ldots n\}$, then the CDF of the combined distribution x_c is given by the below equation.

The generation of adjusted redistributions and their combination was achieved by employing Monte Carlo methods to generate a sufficiently large random sample based on the parameters of the gamma distribution, readjustment of the sample, calculation of cumulative probabilities and an interval based calculation of geometric means across transformed CDFs. The results of the TBF prediction before and after the ensemble process are shown in Figure 9.

Since the current model updates the TBF prediction only when new information in the form of an ATA message event occurs, sparse event counts in the latter half of the dataset are seen to affect model performance. However,

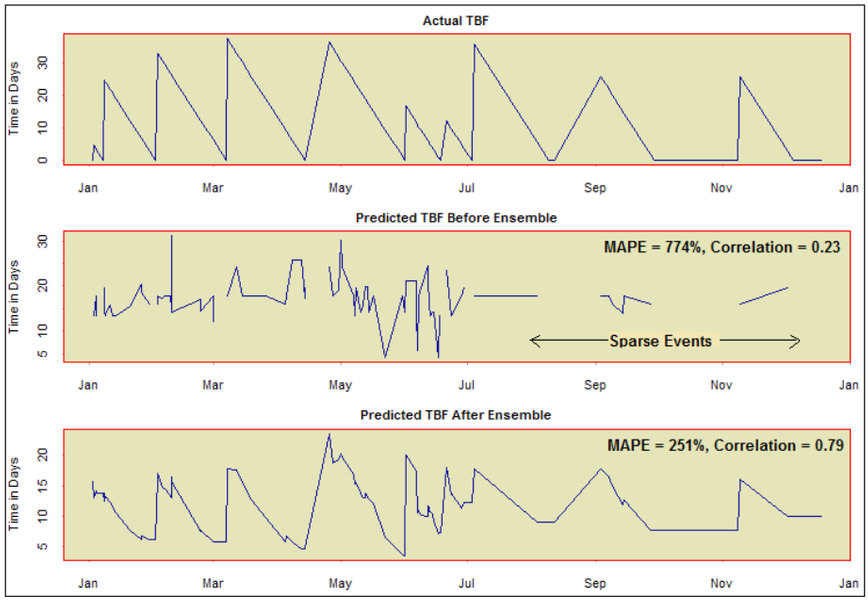


Figure 9: Visualization of Results for a specific Aircraft (Top) Actual TBF (Middle) TBF from Regression Tree Model (Bottom) TBF from Stochastic Ensemble

we observe that the Stochastic Ensemble of TBF distributions not only models the progressive deterioration and approximates the saw-tooth waveform observed in the actual TBF, it is also able to reduce the Mean Average Percentage Error (MAPE) by nearly 70% and improves the correlation by more than 300%, as compared to a standard regression tree model.

CONCLUSION

In this paper we examined the results provided by exploratory data analysis of aircraft health and failure data and used the insights for developing a predictive model for aircraft time to failure. We also explored the use of Hidden Markov Models for delay prediction and conclude that more appropriate training sequences can improve the HMM accuracy of predicting the delay state.

Finally, in this paper, we demonstrated how time adjusted probability distributions can be used with ensemble methods to effectively leverage stochastic information and model progressive deterioration of aircraft health. The approach also significantly improves prediction accuracy and correlation by combining predictions from several models.

REFERENCES

https://en.wikipedia.org/wiki/ATA_100 (Wikipedia ATA list)

Létourneau, S., Fazel, F., & Matwin, S. (1999). *Data Mining Based Component replacement*. National Research Council Cananda.

Salfner, F. (2005). Predicting failures with hidden Markov models. In *Proceedings of 5th European Dependable Computing Conference (EDCC-5)*, pages 41–46, Budapest, Hungary, Apr. 2005. Student forum volume.

Rabiner, L. R., & Juang, B. H. (1986). *An Introduction to Hidden Markov Models.* IEEE ASSP Magazine.

Article can be accessed online at http://www.publishingindia.com

Is Your New Product Really Boosting our Sales? An Econometric Model to Quantify the Cannibalization Effect

Vamse Goutam*

Introduction

As a manufacturer introduces a new product to compete in the market, it is competing with the already existing products of the same manufacturer. Cannibalization is an extremely important concept in marketing – a decrease in sales, revenue or market share of an already existing product of the category due to the introduction of a new product. Most top manufacturers use their already existing brand names to circumvent if not all but some barriers for entry of a new product as a part of line/brand extension strategy, believing that an already successful brand would augment the new product's sales. But, gaining movement at the expense of the parent product is not a real success. A new product is considered successful when it would have created its own market either by creating new customer base (market expansion) or through shift of customers from the competitor's base (customer switching from the competitors brand to the new product). But, there is another possibility as well – the new product might actually intrude into the existing market of the older products, thus eating away the sales of the parent product, in such a case the sales incurred by the new product is not a true incremental sales. In fact it is very critical to acknowledge that for most of the new products some amount of cannibalization is inevitable. However one could take initiatives to keep the cannibalization as minimal as possible, but to expect no cannibalization would be utopian.

Though, cannibalization has an inherent negativity, but when carefully planned and executed, it could be extremely effective, by ultimately growing the market. For example, in the chocolate market in Britain, the decline in Kit Kat's share was arrested by the launch of a new chunkier bar, which undoubtedly cannibalized the market for the original. Firms, also intentionally cannibalize their own products by producing marginally improved products, this is more common with mobile industry. The idea is to persuade the customers to buy a new upgraded version which is always expensive than the original.

Thus, understanding and quantifying cannibalization becomes a very important factor in measuring the success rate of a new product. This paper concentrates on how one could leverage econometrics in the field of marketing to quantify the cannibalization effect of a new product on the already existing products and the portfolio of the same manufacturer.

For the purpose of quantifying the cannibalization impact, we use the SUR (seemingly un-related models) which allows us to account for the fact that the sales of a particular product within a store will not just depend on its price and its own other features but will also depend on the interplay effects between the target and the other products within a store. SUR is a system of equations with multiple equations instead of one single equation with each equation representing an individual business relationship. These individual equations can be estimated with OLS as well but what makes SUR more efficient is that SUR lets the individual equations to be linked and interact with each other. A popular example of SUR is demand estimation for example:

$$Y_{pepsi,t} = X_{pepsi,t}\beta_{pepsi} + E_{pepsi,t}$$

* Bangalore, India Vamse.kobe09@gmail.com

$Y_{coke,t}=X_{coke,t}\,\beta coke+E_{coke,t.}$

Where Ypepsi is the quantity demand (sales) for Pepsi, X pepsi represents vector of regressors such as price of Pepsi, its promotion etc. and "t" represents time index.

The approach used in this paper gives us an absolute sales that was cannibalized by new product from its parent brand or from the portfolio as a whole. This enables the manufacturer to understand the percentage contribution by the new product in the category growth in terms of incrementality and thus one would know if the new innovation has actually fetched the desired results or not (its success rate).

Methodology and Model Form

Seemingly Unrelated Regression (SUR) is a system of linear equations that is not simultaneous. The concept was first considered by Zellner in 1962. The conceptual idea behind SUR is that each and every individual linear equation in the system is conceptually related to one another and aside this conceptual relationship the equations do not hold any outward relationships. Each equation in the SUR system models a different dependent variable and the vector of regressors need not to be same. For instance consider the below system of equation with three individual equations (Three dependent variables are considered as a group because they bear a conceptual relationship, but relationship, but other than this conceptual relationship the variables hold no connection):

$$Y_{t,1}=X_{t,1}\,\beta_1+U_{t,1} \quad (1)$$

$$Y_{t,2}=X_{t,2}\,\beta_2+U_{t,2} \quad (2)$$

$$Y_{t,3}=X_{t,3}\,\beta_2+U_{t,3} \quad (3)$$

At first glance the above three equations seem unrelated - each individual equation has different vector of regressors and different dependent variables. Hence one would typically claim to have them estimated separately. Of course if the three equations are unrelated then they need to be estimated separately, but if these three equations hold a relationship that is not explicitly controlled by the regressors but brought forward by the their error terms ($U_{t,i}$) then it would be more efficient to have all three estimated jointly. The idea is better understood for a single equation with serial correlation, if $U_{t,1}$ is correlated with $U_{t,2}$ and $U_{t,2}$ with $U_{t,3}$ then the knowledge of $U_{t,2}$ and $U_{t,3}$ can help us in better accuracy. Thus, if the error terms of the three equations are correlated (estimated variance covariance matrix becomes non – singular) then we would have a more efficient estimator by estimating all the equations jointly.

SUR is a two-step generalized least squares (GLS) estimator for the model where the estimated variance covariance matrix of the error terms in the system is non-singularity. On many applications in economics, business and management, the assumption of non-singularity may be violated and that is also the case with estimating cannibalization.

In order to quantify the cannibalization impact of a new product over the already existing products one needs to consider the fact that the sales of a particular product within a store will not just depend on its price and its own other features but will also depend on the interplay effect between the target and the other products within a store and SUR allows to account for this fact.

A large US confectionary manufacturer has introduced a new product differentiating it by a new flavor and would like to understand the cannibalization impact of the new product over the already existing products in the portfolio. Thus, wanting to understand if the sales incurred by the new product was actually an incremental sale or cannibalized from its parent products.

The portfolio consists of four products excluding the new product. For simplicity we call them product1, product2, product3, product4 and the new product. Thus, we would want to understand what percentage of the new product sales was actually rooted out from the already existing four products in the portfolio and what percentage of the new products sales was a true incremental sale.

Prior to the SUR system framework it is better to look at one single equation to get a better understanding on the model form. Assuming that we are looking at only one of the products in the portfolio – product 1. The equation will take the below form:

$$S_1=\alpha(Calculated\ \ Price1)\beta_1(PI_1)\beta_2 e\beta_3(Feature1)+\beta_4(Display)+\beta_5(Feat\&Disp)+\beta_6(Growthterm)$$

1 1

Where,

S1 is the sales of product1.

Calculated price of product1 is an indication of what the price points in the store be in the absence of price discounts. The whole idea of calculated price is to see whether the price discounts are effective or not.

PI1 (Price Index) is a ratio of actual price (this will always be lesser than or equal to the calculated price of the product) observed in the retail store for product1 to calculated price of product1 – actual price is price after discounts. Thus, as the PI increases the gap between the calculated price and the actual price reduces which means lesser discounts at the store hence lesser sales indicating that the actual price index has a negative relationship with the sales.

Feature, Display and Feature and Display are dummy variables capturing the promotional activities with an objective to boost the sales of product1, for simplicity we refer feature, display and feature and display as the promotional variables. Since the investment in the promotional variables is made with an expectation to boost the sales over and above the base sale, we expect these variables in the model to show a positive relationship with the sales of the product.

Growth term is the variable of interest here, this is the variable that captures the impact of the cannibalization – the impact of a new product being introduced on the already existing products and the βs are the coefficients to be estimated. The growth term is defined as a ratio of the new product sales to the portfolio sales as a whole. In simple words the variable represents share of the new product in the portfolio.

The above model is in a multiplicative form such that it allows for the interactions between variables in order to account for the fact that the sales of a product is not determined only by its standalone features (Price and other promotional variables) but also that these features interact with each other to incur sales. The above multiplicative model is transformed in order to get an estimable form – we double - log linearize the above multiplicative model and the below is the new model form:

$$ln(S_1) = ln\alpha + \beta_1 ln(Calculated\ Price_1) + \beta_2 ln(PI_1) + \beta_3 Feature_1 + \ldots + \beta_6 Growth\ Term_1 + \varepsilon_1$$

With four already existing products in the portfolio, the SUR system will consist of four independent equations like the above each modeling one of the four already existing products in the portfolio against its own price points, promotions and the growth term. The model form will look as below:

$$ln(S_{1,t}) = ln\alpha_1 + \beta_1 ln(CalculatedPrice_{1,t}) + \beta_2 ln(PI_1) + \beta_3 Feature_{1,t} + \ldots\ldots + \beta_6 GrowthTerm + \varepsilon_{1,t} \quad \text{-- (1)}$$

$$ln(S_{2,t}) = ln\alpha_2 + \beta_1 ln(CalculatedPrice_{2,t}) + \beta_2 ln(PI_2) + \beta_3 Feature_{2,t} + \ldots\ldots + \beta_6 GrowthTerm + \varepsilon_{2,t} \quad \text{-- (2)}$$

$$ln(S_{3,t}) = ln\alpha_3 + \beta_1 ln(CalculatedPrice_{3,t}) + \beta_2 ln(PI_3) + \beta_3 Feature_{3,t} + \ldots\ldots + \beta_6 GrowthTerm + \varepsilon_{3,t} \quad \text{-- (3)}$$

$$ln(S_{4,t}) = ln\alpha_4 + \beta_1 ln(CalculatedPrice_{4,t}) + \beta_2 ln(PI_4) + \beta_3 Feature_{4,t} + \ldots\ldots + \beta_6 GrowthTerm + \varepsilon_{4,t} \quad \text{-- (4)}$$

The SUR system will now have four independent double–log linearize equations with the same growth term repeated in all the equations.

Data

Data is a cross sectional data that has been simulated for1000 stores over a period of 78 weeks. This is a point of sales data where the sales information along with the data on the independent variables have been simulated for each and every individual store for each week. The 78week time period is actually divided into two segments, 52 weeks prior the new product launch and 26 weeks post the launch. The idea here is to have at least one whole year of data on the already existing products in the portfolio, sothat the model has enough data points to learn and capturethe pattern of these products and also give enough timefor the new product to stabilize in the market.

Start Period:02/22/2014

End Period: 08/15/2015

Total # of Stores:1000

Total # of Weeks:78

Total # of Observations:78,000

Initial Diagnosis

High promotional support (Display and Price Discounts)

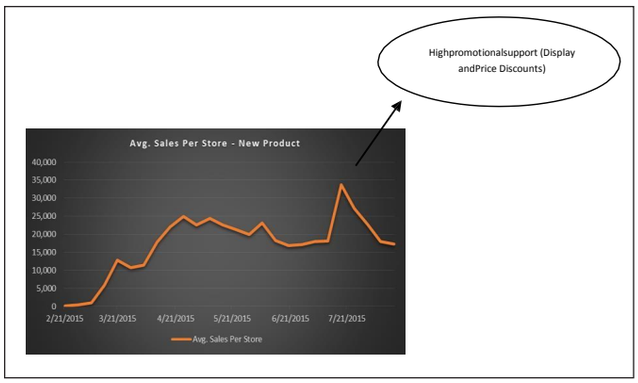


Figure 1.1: (New Product Sales per Store by Week)

Figure 1.2: (Prodcut 1 Sales per Store by Week)

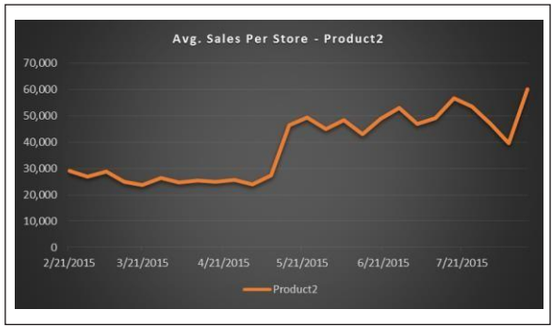


Figure 1.3:

Figure 1.4:

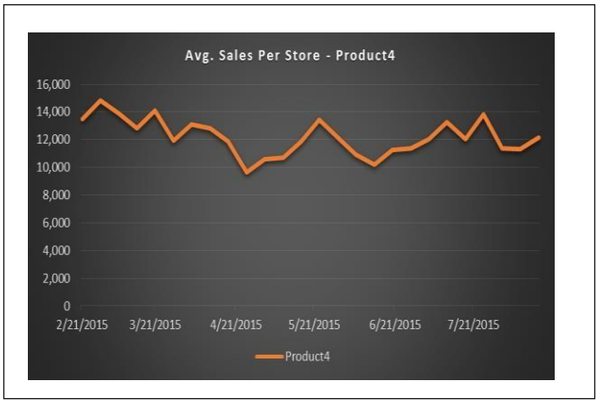


Figure 1.5. (Product 4 Sales per Store by Week)

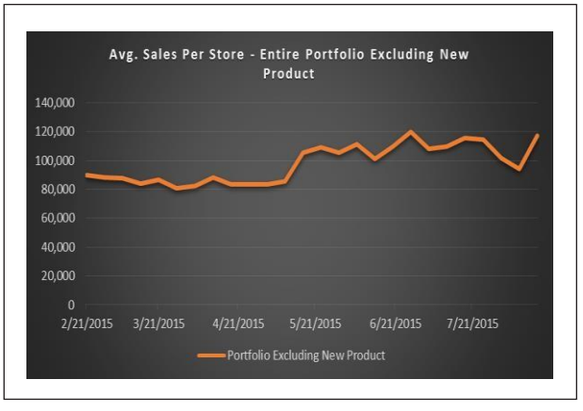


Figure 1.6. (Entire Portfolio Sales per Store Excluding New Product)

Looking at the initial diagnosis (Sales Per Store by Week) the first thing that one could claim is that – "distributional gains and merchandising support drove the sales of the new product – see Figure 1.2", there has been a combination of promotional activity (In – Store Display and 30% price discount) incurred during the stabilization period, which

supported the launch of the new product.

Looking at the plots of the already existing products in comparison to the new product one could claim that the cannibalization effect within the portfolio has been really contained to minimal and this could have been carried out with appropriate merchandising support by the manufacturer for the already existing products in the portfolio.

The trend of the already existing products in the portfolio is either neutral or upward moving trend except product 4 – see Figure 1.5, which shows a slight negative sloping trend, but it is too early to conclude that this downward sloping trend is indeed a cannibalization effect by the new product, however with the above plots it is now safe to claim that the cannibalization has been really minimal. Figure 1.6 – Sales per store indicates that even at the portfolio level there has been almost no cannibalization effect, even if it had incurred it has been contained to very minimal. This is even more evident with the below plot – Figure 1.7 –

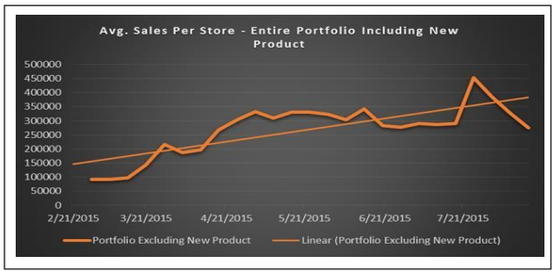


Figure1.7: (Entire Portfolio Sales per Store by Week Including New Product)

If Figure 1.6 is compared with Figure 1.7, one could claim that the new product launch has actually been incremental to the portfolio and indeed incurred incremental sales.

For modeling SAS was used and procedure PROC SYSYLIN was implemented, below is a

sample code of procedure PROC SYSLIN:

```
procsyslin data = Model_Datasuroutest=est_model
out=pred_model;

model Sales1 = cal_price1 PI2 display1 feature1 feat_
disp1 growth_term seasonality trend;

model Sales2 = cal_price2 PI2 display2 feature2 feat_
disp2 growth_term seasonality trend;

model Sales3 = cal_price3 PI3 display3 feature3 feat_
disp3 growth_term seasonality trend;

model Sales4 = cal_price4 PI4 display4 feature4 feat_
disp4 growth_term seasonality trend;

run;
```

In the above code there are four model statements one for each product in the portfolio except the new product. Below is the list of significant variables of the model results for all four equations, one thing to note here is that the below results are output from seemingly unrelated regression estimation. PROC SYSLIN provides both the outputs the OLS and the SUR but the below results are only SUR as they are the ones we are concerned with:

Product1		
Variable	*Estimates*	*Pr>|t|*
Intercept	(0.024)	0.030
Price	(0.405)	0.025
PI	(0.003)	<.0001
Seasonality	0.187	0.023
CanibllizationEffect	(0.044)	<.0001
Promo1(Display)	0.114	0.102
Promo2(Feature)	0.127	0.003
TREND	0.094	<.0001

Product2		
Variable	*Estimates*	*Pr>|t|*
Intercept	(0.113)	<.0001
Price	(0.188)	0.002
PI	(0.470)	<.0001
Seasonality	0.286	0.183
CanibllizationEffect	(0.099)	0.022
Promo1(Display)	0.035	<.0001
Promo2(Feature)	0.153	<.0001
Promo3(Featureand	0.337	<.0001
TREND	0.039	<.0001

Product3		
Variable	*Estimates*	*Pr>|t|*
Intercept	0.044	<.0001
Price	(0.592)	0.002
PI	(0.518)	<.0001
CanibllizationEffect	(0.076)	0.022

Product3		
Variable	*Estimates*	*Pr>|t|*
Promo1(Display)	0.346	<.0001
TREND	0.439	<.0001

Product4		
Variable	*Estimates*	*Pr>|t|*
Intercept	0.242	0.030
Price	(0.432)	0.025
PI	(1.456)	<.0001
Seasonality	0.187	0.023
CanibllizationEffect	(0.380)	<.0001
Promo1 (Display)	0.075	0.102
Promo2 (Feature)	0.266	0.003
TREND	0.998	<.0001

Across models we see that the cannibalization term has become significant and negative. This indicates that there has been some amount of cannibalization effect because of the new product.

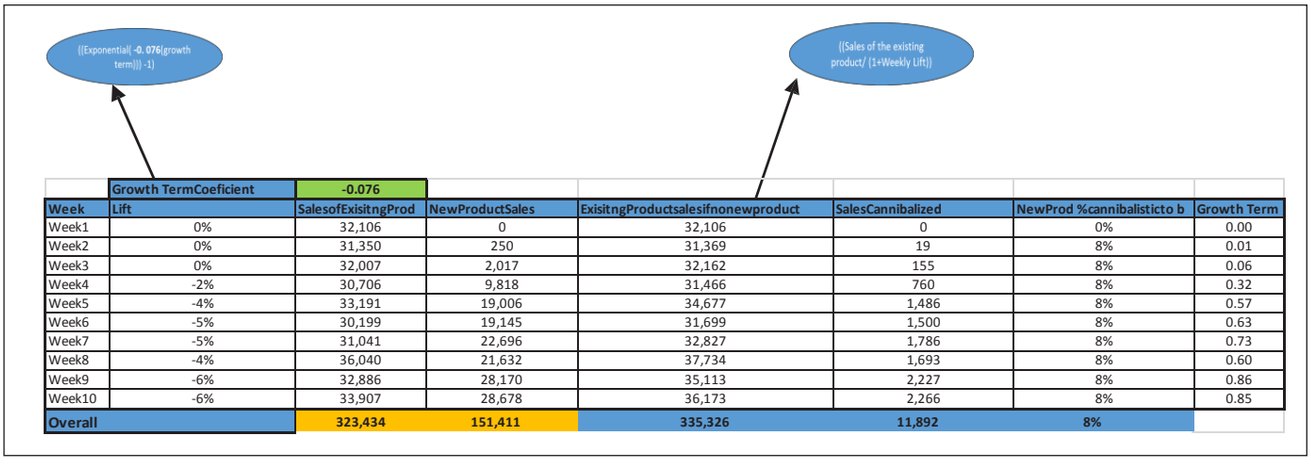
As the above results are from a linear model form and in order to quantify the exact amount of the cannibalized sales and the incremental sales of the new product we need to use the above results and roll them back to their real space i.e. multiplicative model form.

We use the above cannibalization effect estimates and calculate the actual cannibalized sales by the new product for all the 4 products individually and then for the entire portfolio. Due to confidentiality of the data with the client, the entire calculation is not shown in this paper using the entire data, however in the appendix using product 3 a sample calculation is shown, so that the reader gets an idea on how the quantification is done after the modeling stage.

Insights and Actions

- The New product launch has been successful in generating portfolio incrementality, hence driving the brand and the portfolio growth.
- Focus on building strong distribution in early time periods along with merchandising support (evident from Figure 1.1) as it enables strong trial to allow for repeat purchases within the first year of launch.
- Of all the four products the new product tends to have cannibalized product4 the most with 12.13% decrease in its sales. Hence, the downward trend that was observed during the initial diagnosis can now be attributed to the new product launch.
- The overall % of cannibalized sales that the new product accounted is 23% with 77% incremental.
- The 23% of the cannibalization by the new product is distributed as 1.04%, 6.76%, 2.82% and 12.13% for product 1, product 2, product 3 and product 4 respectively.
- Cannibalization sample calculation:
- Assume that fora particular product the growth term estimated through PROC SYSLIN is**-0.076** and we are working with onlyten weeks of data.
- Given that, the original multiplicative model has been linearized, thereforethe estimates needs to berolled back to theoriginal space. Below is given theformulathat we useto roll back the growth term
- *((Exponential(**-0. 076**(growth term))) -1)*– This is calculated for each and every single week. Please look at the below table for further explanation.

Please note the numbers in the above table are only for explanatory, hence were simulated. These numbers were not used as part of the any project that has been executed.



	Growth TermCoeficient	-0.076					
Week	Lift	SalesofExisitngProd	NewProductSales	ExisitngProductsalesifnonewproduct	SalesCannibalized	NewProd %cannibalisticto b	Growth Term
Week1	0%	32,106	0	32,106	0	0%	0.00
Week2	0%	31,350	250	31,369	19	8%	0.01
Week3	0%	32,007	2,017	32,162	155	8%	0.06
Week4	-2%	30,706	9,818	31,466	760	8%	0.32
Week5	-4%	33,191	19,006	34,677	1,486	8%	0.57
Week6	-5%	30,199	19,145	31,699	1,500	8%	0.63
Week7	-5%	31,041	22,696	32,827	1,786	8%	0.73
Week8	-4%	36,040	21,632	37,734	1,693	8%	0.60
Week9	-6%	32,886	28,170	35,113	2,227	8%	0.86
Week10	-6%	33,907	28,678	36,173	2,266	8%	0.85
Overall		323,434	151,411	335,326	11,892	8%	

In the above table the cell highlighted in green is the growth term estimate that was obtained through SUR model.

Column 1 (Week)-states the week number,

Column 2 (Weekly Lift)-states the percentage of cannibalization effect– This is calculated using the below formula

((Exponential(-0.076)*(growth term))– 1),

Column 3 -Sales ofthe Existing product, Column 4 – Sales of the new product launched,

Column 5 – Existing Product sales if the new product has not been launched– This is calculated using the below formula

((Sales of the existing product/ (1+Weekly Lift))

Column 6 – Sales Cannibalized– This is the difference betweenexisting product sales (column 3)

andexisting product sales if the new product hasnot been launched (column 5),

Column 7 & Column 8 - Percentage cannibalization incurred by the new product in the above example it is 8% and growth term (independent variable that actually went into the model).

In the given example, we have estimated the cannibalization sales to be 8% then the incremental sales of the new product would be 100% - 8% = 92%.

Similarly the cannibalization effect is calculated for all the products in the portfolio individually and then the cannibalized sales from all the existing products is summed up across products/equations in the system to arrive at the overall cannibalized sales within the portfolio.

Guidelines for Authors

International Journal of Business Analytics and Intelligence welcomes original manuscripts from academic researchers and business practitioners on the topics related to descriptive, predictive and prescriptive analytics in business. The authors are also encouraged to submit perspectives and commentaries on business analytics, cases on managerial applications of analytics, book reviews, published-research paper reviews and analytics software reviews based on below mentioned guidelines:

Journal follows online submission for peer review process. Authors are required to submit manuscript online at http://manuscript.publishingindia.com

Title: Title should not exceed more than 12 Words

Abstract: The abstract should be limited to 150 to 250 words. It should state research objective(s), research methods used, findings, managerial implications and original contribution to the existing body of knowledge

Keywords: Includes 4–8 primary keywords which represent the topic of the manuscript

Main Text: Text should be within 4000-7000 words Authors' identifying information should not appear anywhere within the main document file. Please do not add any headers/footers on each page except page number. Headings should be text only (not numbered).

Primary Heading: Centered, capitalized, and italicized.

Seconday Heading: Left justified with title-style capitalization (first letter of each word) and italicized.

Tertiary Heading: Left justified and indented with sentence-style capitalization (first word only) in italics.

Equations: Equations should be centered on the page. If equations are numbered, type the number in parentheses flush with the left margin. Please avoid using Equation Editor for simple in-line mathematical copy, symbols, and equations. Type these in Word instead, using the "Symbol" function when necessary.

References: References begin on a separate page at the end of paper and arranged alphabetically by the first author's last name. Only references cited within the text are included. The list should include only work the author/s has cited. The authors should strictly follow APA style developed by American Psychological Association available at American Psychological Association. (2009). Publication manual of the American Psychological Association (6th Ed.). Washington, DC.

Style Check

To make the copyediting process more efficient, we ask that you please make sure your manuscript conforms to the following style points:

Make sure the text throughout the paper is 12-point font, double-spaced. This also applies to references.

Do not italicize equations, Greek characters, R-square, and so forth. Italics are only used on p-values.

Do not use Equation Editor for simple math functions, Greek characters, etc. Instead, use the Symbol font for special characters.

Place tables and figures within the text with titles above the tables and figures. Do not place them sequentially at the end of the text. Tables and figures must also be provided in their original format.

Use of footnotes is not allowed; please include all information in the body of the text.

Permissions

Prior to article submission, authors should obtain all permissions to use any content if it is not originally created by them. When reproducing tables, figures or excerpts from another source, it is expected to obtain the necessary written permission in advance from any third party owners of copyright for the use in print and electronic formats. Authors should not assume that any content which is freely available on the web is free to use. Website should be checked for details of copyright holder(s) to seek permission for resuing the web content

Review Process

Each submitted manuscript is reviewed first by the chief editor and, if it is found relevant to the scope of the journal, editor sends it two independent referees for double blind peer review process. After review, the manuscript will be sent back to authors for minor or major revisions. The final decision about publication of manuscript will be a collective decision based on the recommendations of reviewers and editorial board members

Online Submission Process

Journal follows online submission for peer review process. Authors are required to register themselves at http://manuscript.publishingindia.com prior to submitting the manuscript. This will help authors in keeping track of their submitted research work. Steps for submission is as follows:

1. Log-on to above mentioned URL and register yourself with "International Journal of Business Analytics & Information"
2. Do remember to select yourself as "Author" at the bottom of registration page before submitting.
3. Once registered, log on with your selected Username and Password.
4. Click "New submission" from your account and follow the 5 step submission process.
5. Main document will be uploaded at step 2. Author and Co-author(s) names and affiliation can be mentioned at step 3. Any other file can be uploaded at step 4 of submission process.

Editorial Contact
Dr. Tuhin Chattopadhyay
Email: dr.tuhin.chattopadhyay@gmail.com
Ring: 91-9250674214

Online Manuscript Submission Contact
Puneet Rawal
Email: puneet@publishingindia.com
Ring: 91-9899775880

www.manuscript.publishingindia.com

PublishingIndia.com

Publishing India Group

Plot No. 56, 1st Floor, Deepali Enclave
Pitampura, New Delhi-110034, India
Tel.: 011-47044510, 011-28082485
Email: info@publishingindia.com
Website: www.publishingindia.com

ISSN 2321-1857
01
9 772321 185001

International Journal of Business Analytics and Intelligence

SUBSCRIPTION DETAILS

Dispatch Address:-
The Manager,
International Journal of Business Analytics and Intelligence
Plot No-56, 1st Floor
Deepali Enclave, Pitampura
New Delhi -110034
Ph - 9899775880

Subscription Amount for Year 2016

	Print	Print + Online
Indian Region	Rs 2700	Rs 3400
International	USD 150	USD 180

Price mentioned is for Academic Institutions & Individual. Pricing for Corporate available on request. Price is Subject to change without prior notice.

Payment can be made through D.D./at par cheque in favour of "Publishing India Group" payable at New Delhi and send to above mentioned address.

Disclaimer

The views expressed in the Journal are of authors. Publisher, Editor or Editorial Team cannot be held responsible for errors or any consequences arising from the use of Information contained herein. While care has been taken to ensure the authenticity of the published material, still publisher accepts no responsibility for their accuracy.

Journal Printed at Anvi Composers, Paschim Vihar.

Copyright

Printed and published by Publishing India Group, New Delhi. Any views, comments or suggestions can be addressed to – Coordinator, IJBAI, info@publishingindia.com