4th Healthcare Capability Building Seminar

16 Feb 2017, Mumbai

Application of unsupervised learning and Decision Trees on Big Data to validate Actuarial Product designing assumptions

Leading edge technology. Innovation driven.

Whereisinformationlostindata?Whereisknowledgelostininformation?whereisinsightlostinknowledge?Whereiswisdomlostininsight?

Where is information, lost in data?
Where is knowledge, lost in information?
Where is insight, lost in knowledge?
Where is wisdom, lost in insight?

Inspiration: T.S. Eliot (The Rock)

Content

Application of unsupervised learning and Decision Trees on Big Data to validating Actuarial Product designing assumptions

Source: Celent

Goal

- Competition is forcing Insurers to create, deliver and manage products having greater **balance between price and risk**, than in the past
- Actuarial **Pricing** is **sensitive to** the **new business volume written** and **persistency** thus, they are estimated/projected frequently
- Our goal was one step ahead of ‘projection’. It was to identify the steps that could be taken in case, NBV and persistency targets are not met.
- Our aim was to get answers for following questions:
  - **Why a confirmed lead/soft lead/prospect doesn’t convert to sale?**
  - **Why an insurance policy gets lapsed/not renewed?**
- That is, to understand the ‘Leads’ (sale opportunities), Agents’ performance and Products better so that the company can maximize sale in order to meet NBV and persistency targets

Business Case

- Two companies selling different insurance products (1st one selling L&H, 2nd one P&C/Health) were falling short of initially planned volume.
- Sales growth rate though positive, less than planned causing NBS
- Avg. Conversion Rate was (in single digits) lower than the Industry average in similar category)
- Recent expenses on media and advertisement significantly increased, still leads were dropping.

- Sales process was outsourced to more than one 3rd party provider, less direct control
- Significant difference in Agent's performance, and high turnover of agents
- Sustainability of product, sales process, and actuarial assumptions, all under review!

	(Q4) YOY Trend - Leads -Decreasing		
	13Q4 over 12Q4	14Q4 over 13Q4	15Q4 over 14 Q4
Leads	-1%	5%	-11%

	(Q4) YOY Trend -Sales - Increasing		
	13Q4 over 12Q4	14Q4 over 13Q4	15Q4 over 14Q4
Sales	60%	57%	9%

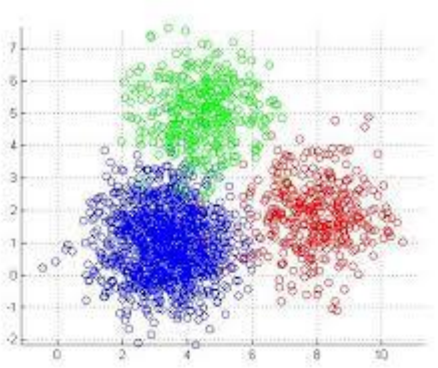
	(Q4) YOY Trend -Leads -Decreasing		
	13Q4 over 12Q4	14Q4 over 13Q4	15Q4 over 14Q4
Max Attempt	-41%	-17%	-12%

Key Words

- **Machine learning**
  - A type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed
  - focuses on the development of computer programs that can change when exposed to new data
- Similar to **data mining as** both systems search through data for patterns. However, instead of extracting data for human comprehension -- as in the case of data mining -- machine learning uses that data to detect patterns in data (another set) and adjust program actions accordingly

- Machine learning algorithms are often categorized **supervised** or **unsupervised**. Supervised algorithms can apply what has been learnt in the past to new data. Unsupervised algorithms can draw inferences from datasets.
- Basic **difference** in layman terms : In **supervised learning**, the output datasets are provided which are used to train the machine and get the desired outputs whereas in **unsupervised learning** no datasets are provided, instead the data is clustered into different classes

Key Words



Unsupervised Clustering : allowing data set to cluster on similar statistics without training input variables to observed outputs which is the case in supervised

Hierarchical clustering: This algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pairwise distance between the data point. Repeating the process again, distance between the data point is recalculated to get one less cluster and so on till we end up with single cluster.

Decision Trees: A decision tree is a very specific type of [probability tree] that enable you to make a decision by discovering patterns in the data that relate data attribute with a target (class) attribute

Big Data: extremely large data sets – structured or unstructured - that may b analysed computationally to reveal patterns, trends, and associations, especiall relating to human behaviour and interactions. Mainstream definition of big data a the three Vs – (1) Volume, (2) Velocity and (3) Variety

SUTHERLAND GLOBAL SERVICES®

Analysis Approach & Challenges

Leading edge technology. Innovation driven.

Analysis Approach

Analytics as usual…

1. Leads Vs Sales trend
2. Sale Conversion analysis - Product Type, Region wise, Month over Month, Age Distribution and Gender
   - Principal components driving Sales
   - Time of Call Vs Sale commencement
   - Call origin sources wise (Media communication and call me later call origin and Customer affiliation with the Company)
3. Cross selling opportunity using Product affinity (both statistical and descriptive support)
4. Lead Leakage Analysis:
   - Reasons for customer driven leakages
   - Call back leakage study hour over hour
   - Not Interested leakage study hour over hour

Analysis Approach

Continued…

5. Agents performance Analysis
   - Distribution of policies sold
   - Region wise stacking
6. Customer profiling
   - Bucketization of Male/Female Policy holders basis other factors (e.g. Age, Region etc.)
   - Identifying suitable customer ages bands – All products and specific products
7. Identifying cluster and layering of policy type sold vs. demographics (a) Supervised (b) Unsupervised – All products and Major products

And then…

8. Customer sentiment analysis – why customers not repeating/prospects not buying

Analysis Approach

Whether it will work? Concerns?

- One of the major concerns – Whether machine learning would work in our systems, with not so perfect data and that too not so largely Big data? Can it replace our learning / experience? Or, whether it will fall short or diverge?
- The **film** studio 20th Century Fox has called in IBM Watson, the supercomputer, to **create** the **trailer** for its upcoming **AI** horror/thriller, Morgan.
- IBM researchers fed Watson more than 100 horror film trailers cut into separate moments and scenes.

- Both the trailers were shown anonymously to a set of people– though, most viewers liked more the one created by human/professionals but, many could not differentiate with confidence
- Positive takeout - 80% of the scenes were common

Where we are headed? The Digital World…

Connected Machines can save up to 150bn across Industries with corresponding increased efficiency

		Segment	Type of Savings	Estimated Value over 15 years*
	Aviation	Commercial	1% fuel saving	$30 billion
	Healthcare	System wide	1% reduction is system inefficiency	$63 billion
	Rail	Freight	1% reduction is system inefficiency	$27 billion
	Power	Gas-fired generation	1% fuel saving	$66 billion
	Oil & Gas	Exploration & Development	1% saving in Capital expenditure	$90 billion

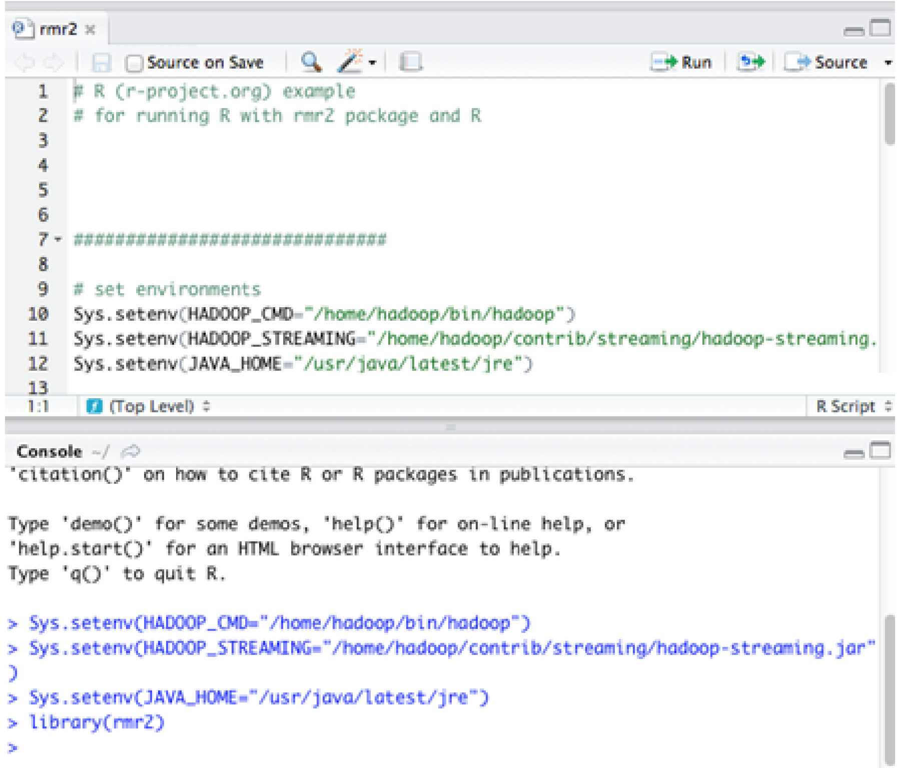
Source: A GE Study

* Billion nominal US Dollar

Methods: Programing Approach

R-Hadoop Ecosystem was created to handle Big Data

- First, R-Hadoop Ecosystem was created. Data collected through various source were compiled, cleaned, aggregated and then condensed using map-reduce function.
- Final analysis was done using R programming




```
dataagg <- function(lines, mjobs=10, rjobs=2) {
  # Map stage
  mlength <- ceiling(length(lines) / mjobs)
  mchunks <- lapply(1:mjobs,
    function(i) lines[(mlength * (i-1) + 1) : (mlength * i)])
  counts <- do.call(c, lapply(mchunks,
    function(x) table(do.call(c, strsplit(clean(x),'\\s')))))

  # Sort
  sorted <- list()
  lapply(1:length(counts), function(i) {
    key <- names(counts)[i]
    if (nchar(key) == 0) return()
    sorted[[key]] <<- c(sorted[[key]], counts[i])
  })

  # Reduce stage
  rlength <- ceiling(length(sorted) / rjobs)
  rchunks <- lapply(1:rjobs, function(i) {
    o <- sorted[(rlength * (i-1) + 1) : (rlength * i)]
    o[!is.na(names(o))]
  })
  rcounts <- lapply(rchunks, function(x) sapply(x,sum))
  out <- do.call(c, rcounts)
  out[order(names(out))]
}
```

```
clust.mr =
 function(data, merge.dataset.size = 10000)
  mapreduce(
   data,
   map =
    function(.,data)
     keyval(1, list(fast. clust(data)[c('n', 'modelName', 'parameters')])),
   reduce =
    function(., models) {
     shrink =
      merge.dataset.size/
      sum(sapply(models, function(m) m$n))
     model =
      fast.mclust(
       do.call(
        rbind,
        lapply(
         models,
         function(m)
          sim(
           modelName = m$modelName,
           parameters = m$parameters,
           n = round(m$n/shrink))[,-1])))
     keyval(
      1,
      list(
       list(
        n = round(model$n*shrink),
        modelName = model$modelName,
        parameters = model$parameters))))})
```

SUTHERLAND GLOBAL SERVICES®

Data & Its Challenges

Leading edge technology. Innovation driven.

Data Used – Size & Sources

- Sample data (of prospects) were drawn from a population of US citizens who were actively/passively looking out for an Insurance policy between 2012-2015
- **Leads Data** – Q4 (2012-2015)– 1,510,196 Leads | 2034k Products sought | 1820 Agents (350 Active)| 167k Customers
- **Sales Data** - Q4 (2012-2015)– 241k Sold Products | 61% Female | 43% Sales in South region
- **Customer Sentiments data** - Q4 (2012-2015) – 632,920 feedbacks, expressions, tagged and untagged conversations, Likes, Emoji etc.
- **Sales channel's CRM data base**, the sold policy details, the interim dispositions, final dispositions, and customer's preference of products
- **Unstructured big data** - collected from various sources such as (a) digitally stored sales channel data, (b) internal records, (c) social media, (d) customer's reactions and (d) insured policy data available in CRM database
- **Derived variables** from available data fields and agents **call recordings**

Data Used – Data Elements and Variables

- **Prospect Data:** (a) Lead start date (b) Lead end date (c) Actual start date (d) Actual end date (e) Planned duration (f) Actual duration (g) Lead generated by (h) Lead originating sources (i) Zip address (j) City (k) State (l) Products preferred.
- **Insured data:** (a) Customer ID (b) Age (c) Gender (d) City (e) Telephone (f) Zip (g) Product Bought.
- **Derived Information:** (a) Trend Base (Month and years) (b) Duration between lead created and end (c) categorized call origin sources (d) Ordering of preferred products (e) Region of lead origin (f) Preferable callback timings (g) Lead created and update hours (h) Number of preferred polices (i) Number of polices bought
- **Data from agents' calls** recorded call when pitching lead: (a) Handling Time (b) number of probing questions asked during conversation (c) split of agent vs prospect talk time
- **From Social Media:** Feedback, expressions, tagged and untagged conversations, Likes, Emoji etc.

Data: Challenges

- **Handling big data** was an obvious challenge - huge volume of data was to be cleansed, new variables from existing fields were to be derived without losing on critical information. Limitation of R in handling such volumes data encouraged us to condense data in R-Hadoop ecosystem, later to be used in chunks in R models.
- **Segmentation of factors** was critical in such large data set. While building decision trees, one needs to be careful that one attribute is not dropped in order to weigh other attributes.
- **Uncertainty in time stamp** of data
- **Multiplicity of sources** and **non conformity** among them. Merging and appending a big task!
- **Unable to order, rank or classify** open ended customer data

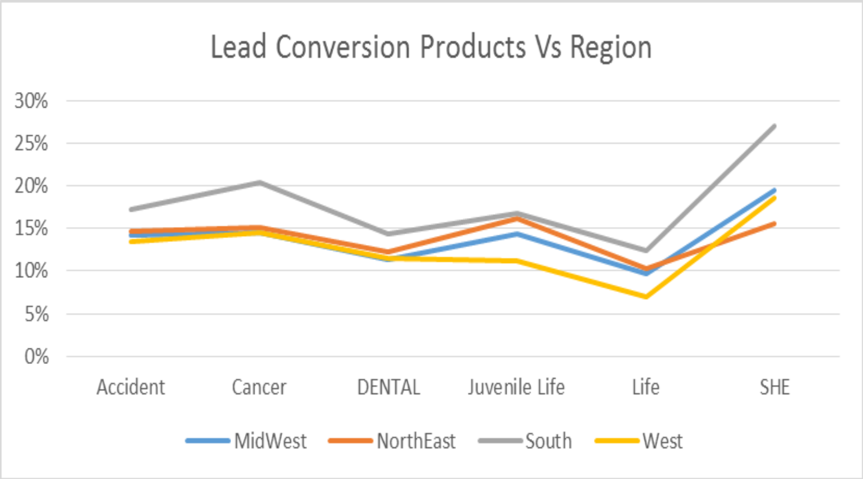
SUTHERLAND GLOBAL SERVICES®

Observations & Findings

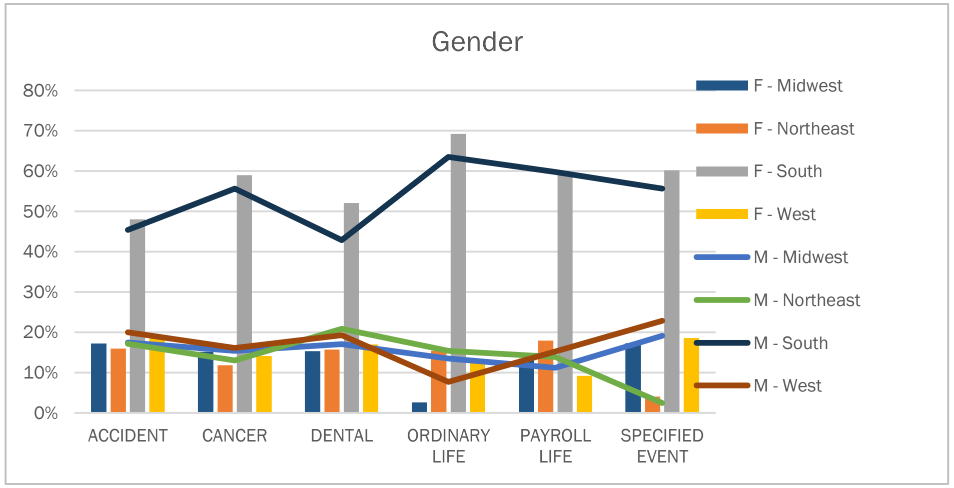
Leading edge technology. Innovation driven.

Region dominance & Gender behavior on Policy Purchase

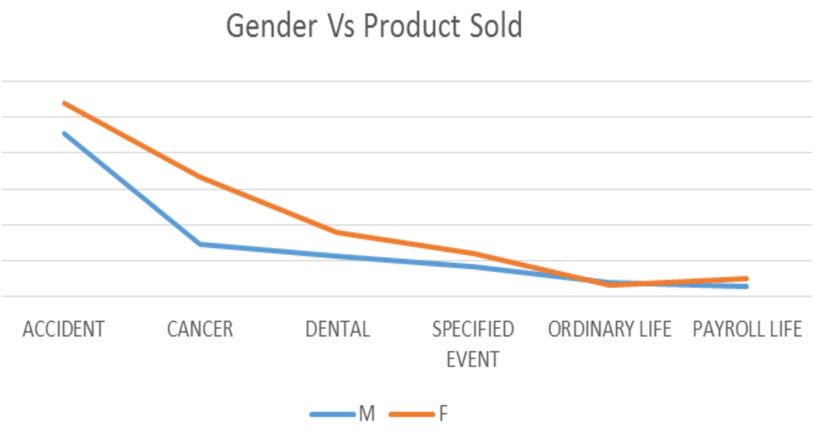
- To assess Region’s impact on policy sales, data was divided in 4 regions (based on zip codes)
- Insignificant variation in conversion rates among regions except, southern region where it was higher than all other three regions across all products
- Overall females bought more policies than males across all products. But, this pattern wasn’t true for all regions. Another proportion matrix representing policies bought by both genders for all regions showed that there were some product types which males tend to buy more than females in different regions

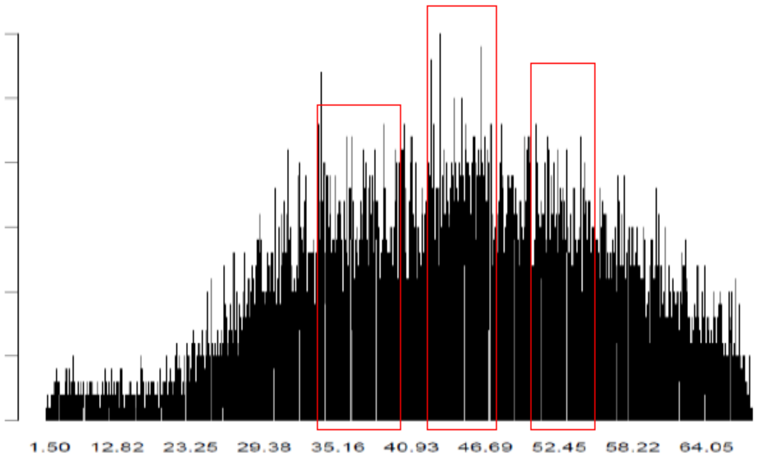


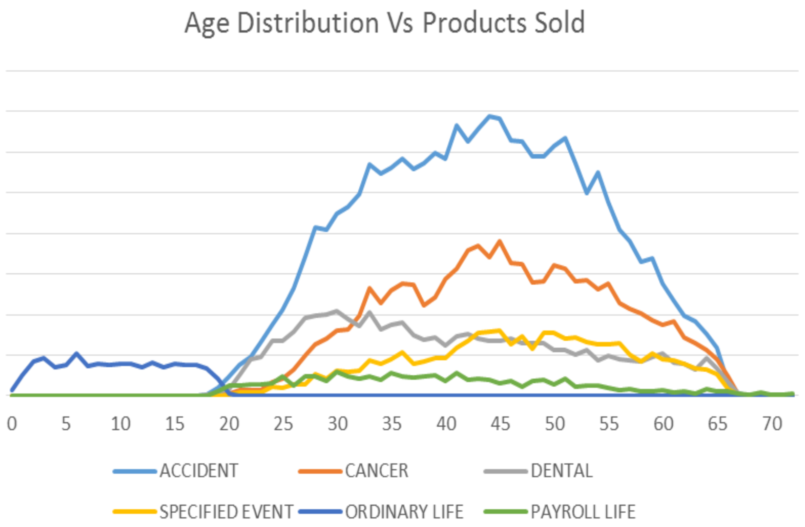
	Difference in Proprtion F to M			
Product Type	Midwest	Northeast	South	West
ACCIDENT	0%	-1%	3%	-1%
CANCER	0%	-1%	3%	-2%
DENTAL	-2%	-5%	9%	-2%
ORDINARY LIFE	-11%	0%	6%	5%
PAYROLL LIFE	2%	4%	0%	-6%
SPECIFIED EVENT	-2%	2%	4%	-4%



Demographic distribution of Policies Sold



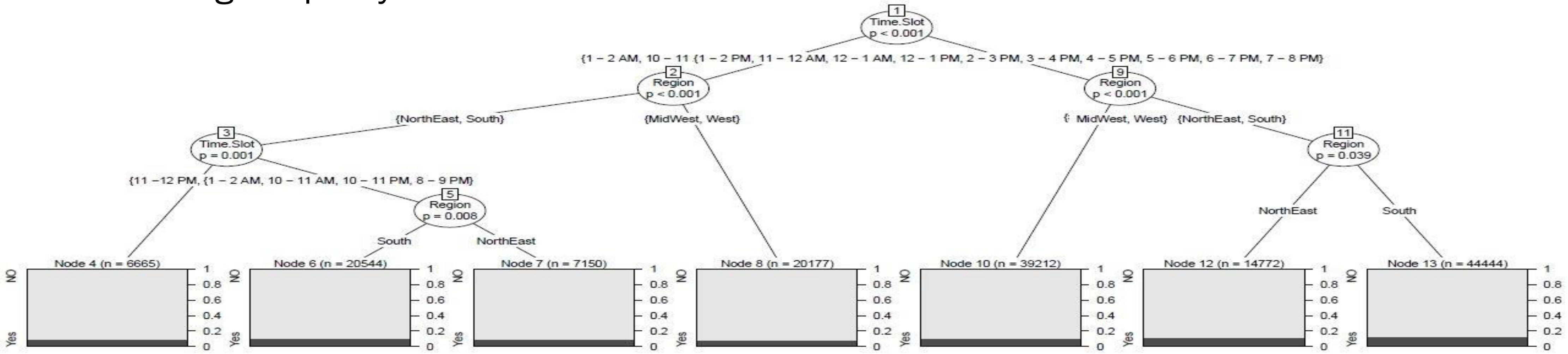




- Plots from machine and descriptive analytivcs, both showed that all products show systematic behavior with peak in age band 40 to 45 whereas, 'Dental' product has peak at 28 to 32.
- Here again, we found something very important i.e. females tend to buy insurance more than males specially 'Cancer' product.

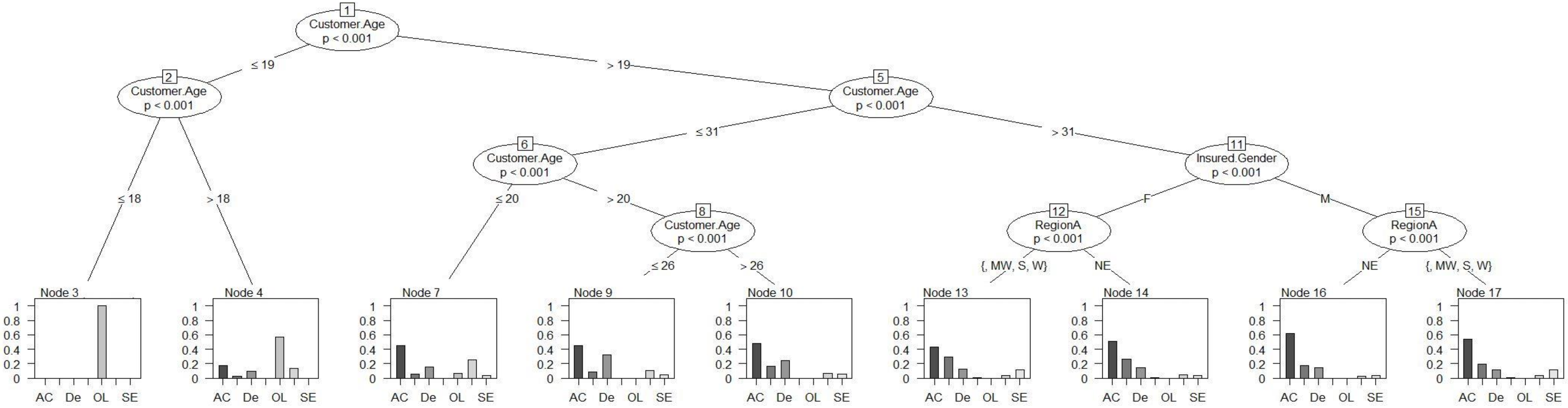
Classification/Decision Tree

- 'Supervised machine learning' technique was applied on classified examples to generate categorical decision tree. The data input in algorithms were labelled as product type. The classifier was learnt by machine from the 'training set' data. Accordingly unseen data or 'test data' was assigned class labels.
- Following decision tree is pruned to represent an snapshot of detailed trees representing buying probability of 'Accident' coverage depending on 'Region' and 'Time' of the day, whence the customer bought a policy.



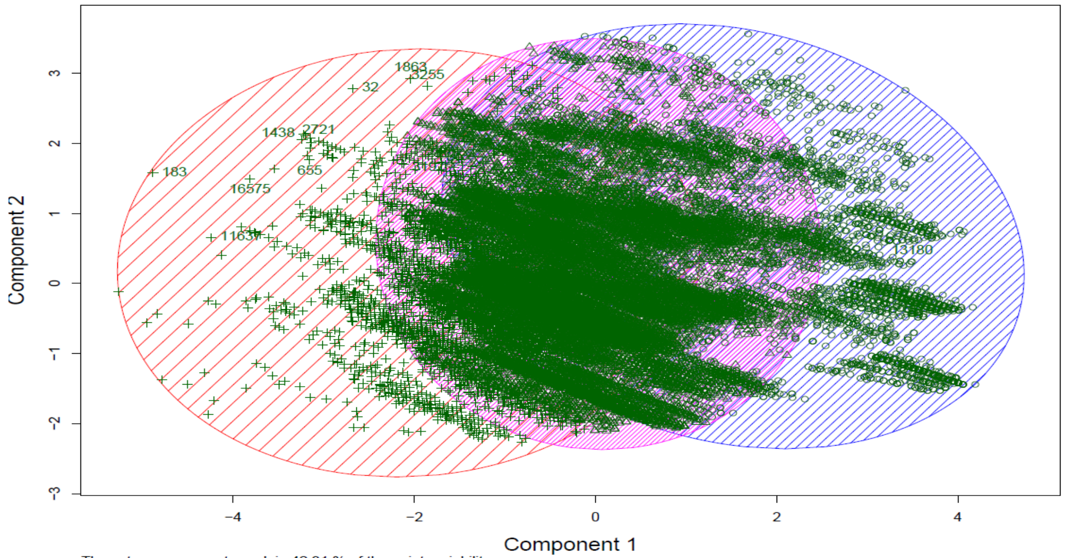
Classification/Decision Tree

- Classifying the customers buying particular products on the basis of demographics, following decision tree got generated. 3 levels of clear nodes (1) Age (2) and (3) Region is visible for all the types of products along with respective probabilities

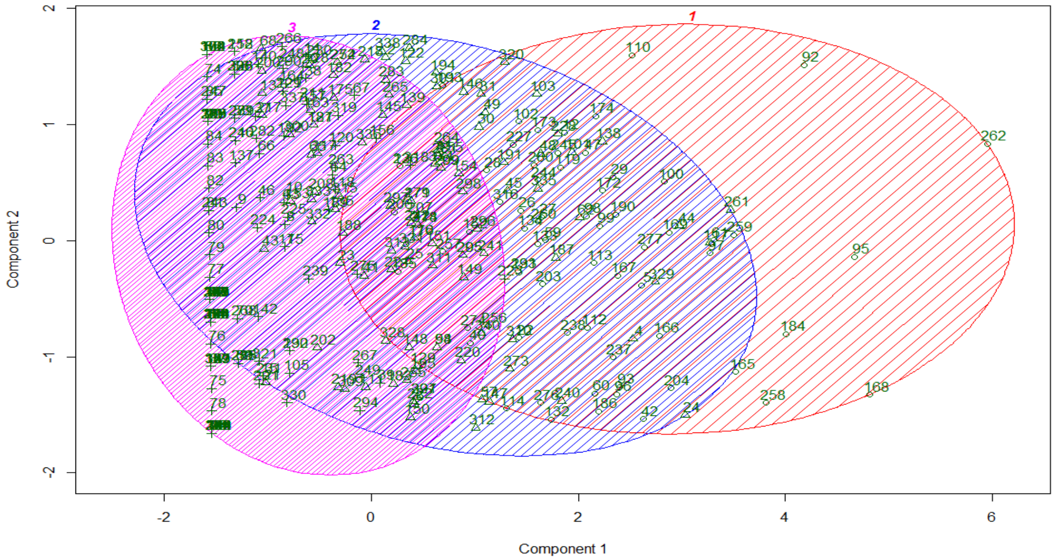


Identifying Clusters

- To identify instances where people buy similar insurance, we applied principal component analysis to subsets of the attribute matrix. Selected principal components captured 42.31% of the variance (left picture). This was used as an input to clustering algorithm instead of the original attribute matrix.
- We then applied principal component analysis to subsets of the attribute matrix after selecting the principal components that captured 66% of the variability (right picture).



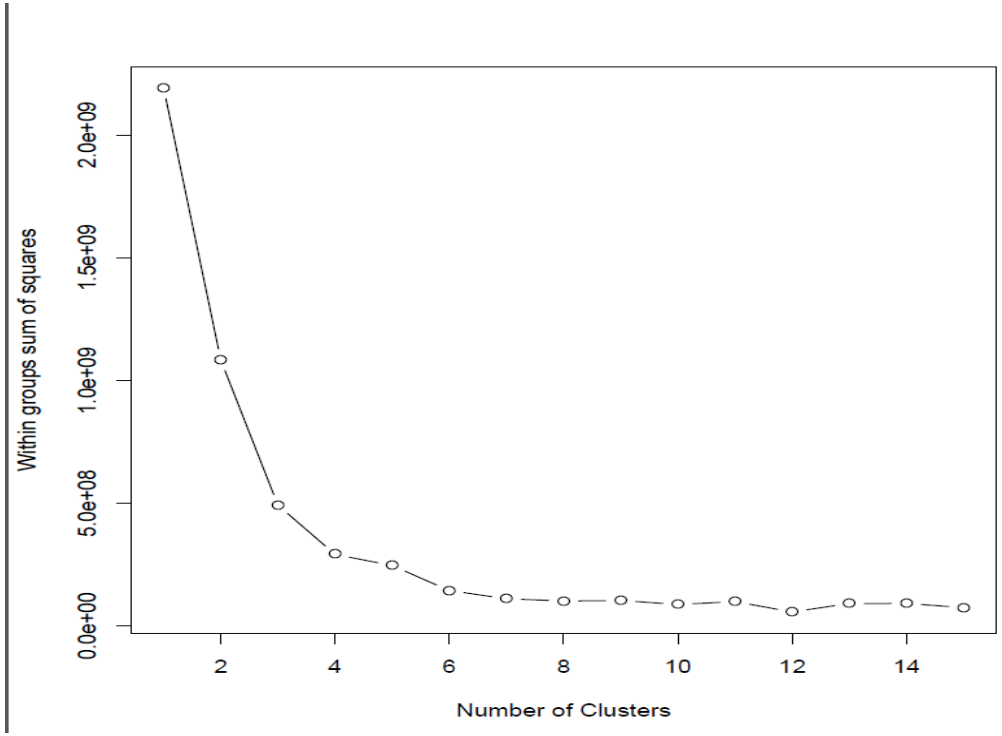
These two components explain 42.31 % of the point variability.

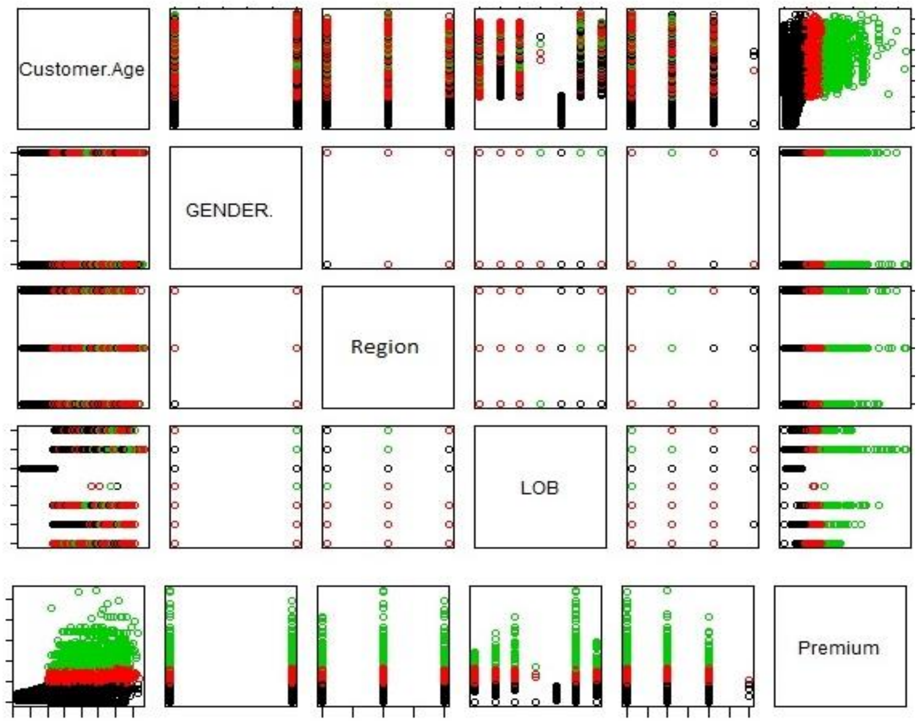


These two components explain 66.16 % of the point variability.

Identifying Clusters

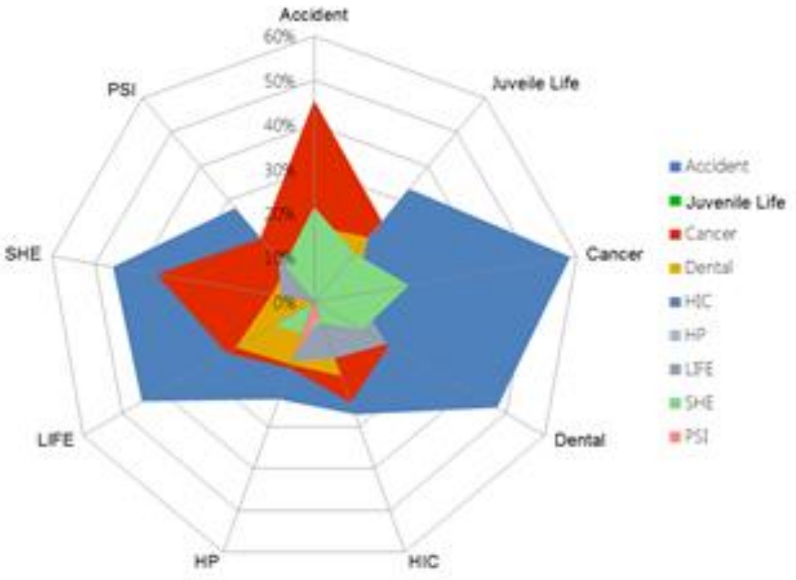
- Used Elbow method to set number of clusters in order to allow for the marginal gain of adding each additional cluster. Largest 'K' with positive marginal gain in our case, was found to be 3
- Descriptive Cluster obtained through Principal Component analysis is placed right below

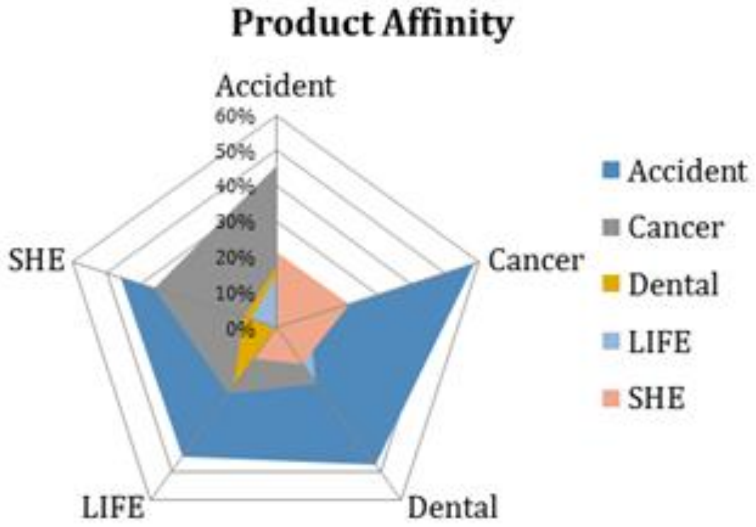


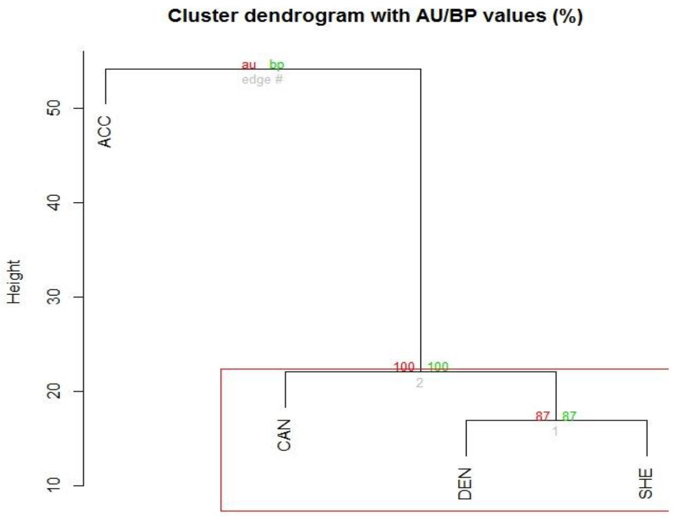


Product Affinity Analysis

- Considering few enquiries/sale of more than 1 product to an individual, we investigated pattern or association between products bought in bundle. 'Accident–Cancer' was found to be most preferred bundle followed by 'Accident-Dental'.
- To ascertain how certain were these trends, we (1) clustered products bought in bundles and then (2) estimated uncertainty using bootstrapping in these clusters through hierarchical clustering and visualization of 'Dendrogram' which gave us Approx. Unbiased probability (AU) as well as Bootstrap probability (BP) as below

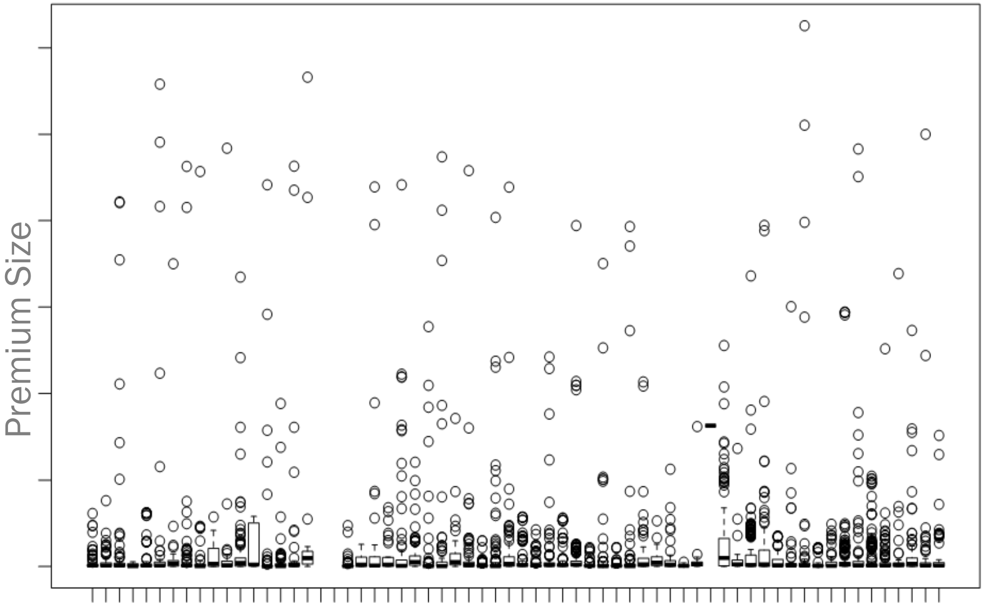


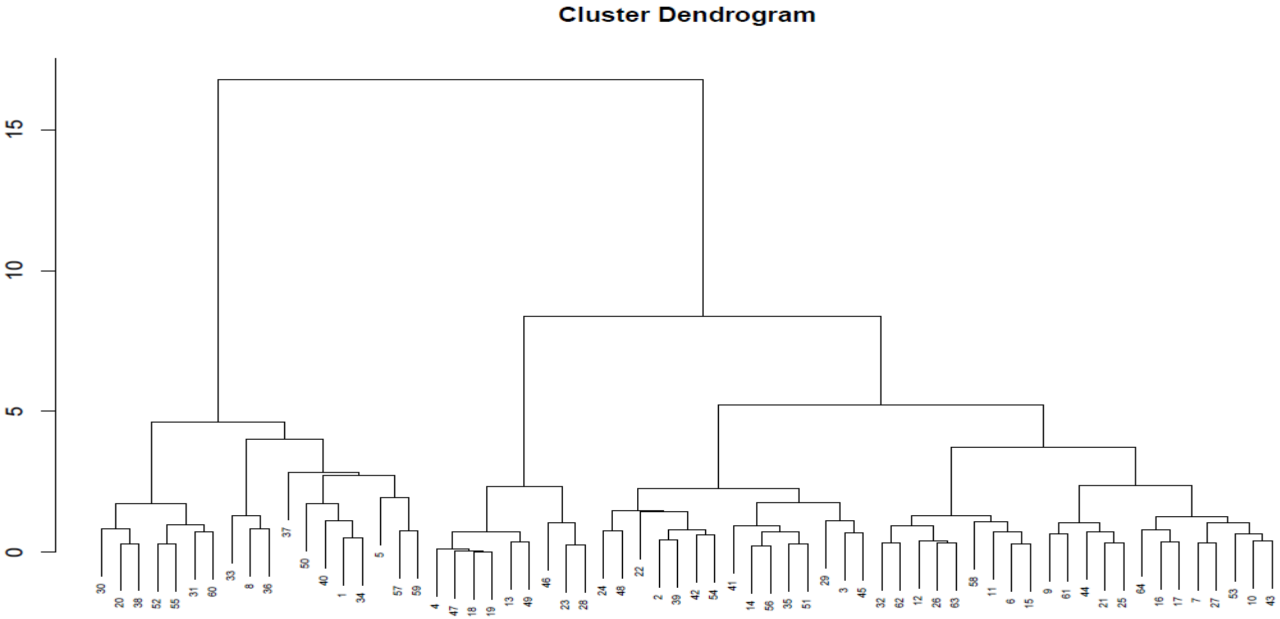




Agent Performance Analysis

- For Agent's performance, earned premium from every single call by different agents was analyzed. Most agents failed to cross $100 premium mark (left plot below). Most agent's attempts are concentrated at $0. The plot also shows the variation in performance, very few performing better than average.
- To investigate further, a decision model was built to cluster of 'performing' and 'non performing' agents together using hierarchical clustering. Resulting Genogram plot is shown below.





Key Findings – Stage 1

1. Leads are dropping ☹, Sales are increasing ☺
2. Conversion rate has improved ☺
3. Though Max attempts reduced still high number of lead got maxed out of attempts (31%) ☹
4. Huge gap between top 10 and bottom 10 performance. Performance varied from 0% to 35%.
5. 42% agent were performing in the range of 0.5% - 5 %. Only 10 % agents were performing very well (29% - 35%)
6. Max sale in Jan, Mar & Oct, least in May & Dec 😐
7. South leads the sale (44%), Others fall b/w 17-22%
8. Accident (50%) is most sold followed by Cancer (22%), Dental (14%), SHE (8%) and Life (4%)
9. Conversion Rates were observed as – Cancer 22%, Acc 17%,Others 10-12%
10. Women bought more (59%) than men (41%). This proportion has not changed since July 2014
11. Probability of a female buying a product is 1.5 times of that of a male buying it
12. Probability of a female buying cancer cover is twice the probability of male buying it
13. Company loyal/associated leads have best conversion rate (32%) followed by Media 26%

Sutherland Global Services®

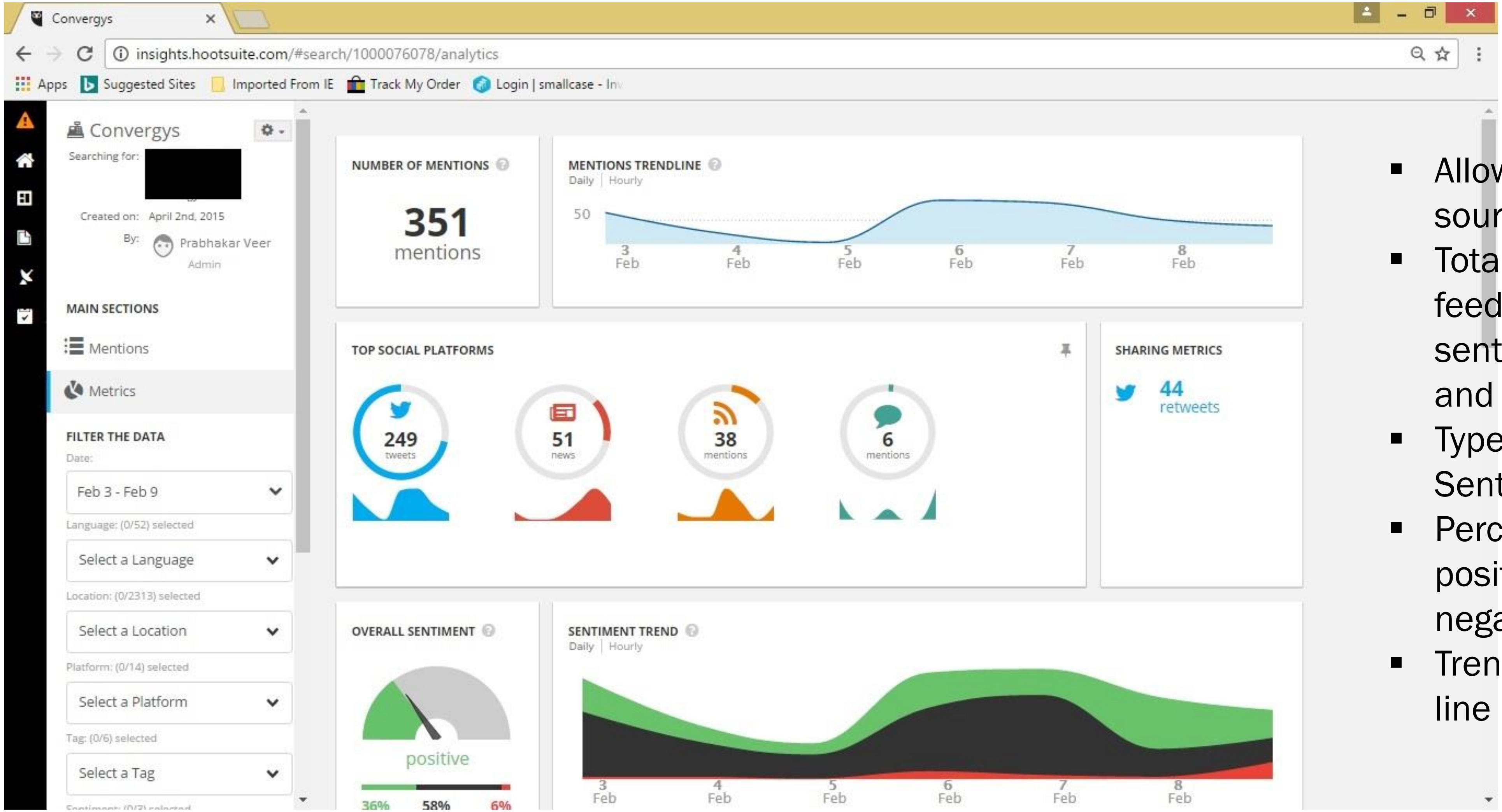
Analysis Approach (stage 2): Got to know what's happening…yet to know Why?

Leading edge technology. Innovation driven.

Sentiment Analysis: Text mining through Web Crawler

Date (GMT)	Platform	Author Name	Author URL	Author Gender	Reach	Country	Region	City	Content	Language	Sentiment	URL	Published (GMT)

Sentiment Analysis: Sample Web Crawler Output



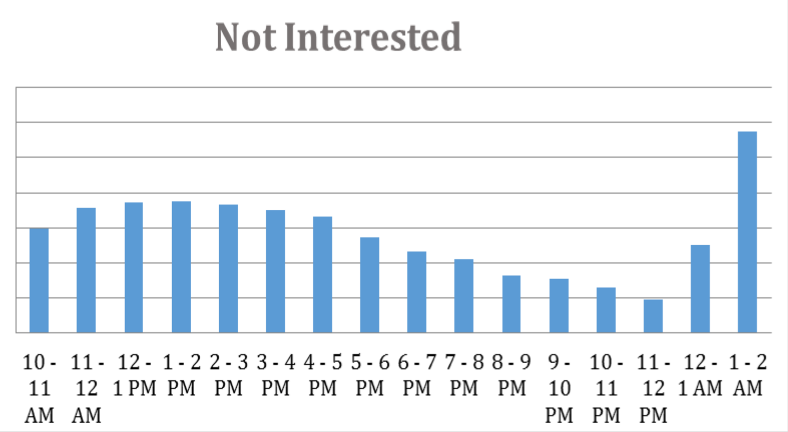
- Allows to know the source
- Total number of feedback/emotions/ sentiments collected and analyzed
- Type of Emotions, Sentiments
- Percentage of positive, neutral and negative feedbacks
- Trend over the time line

Customer Driven Leakage: Call Timing Analysis

Feedbacks can be influenced by customer's availability thus, customer's behavior was analyzed on time scale when approached for policy sale.

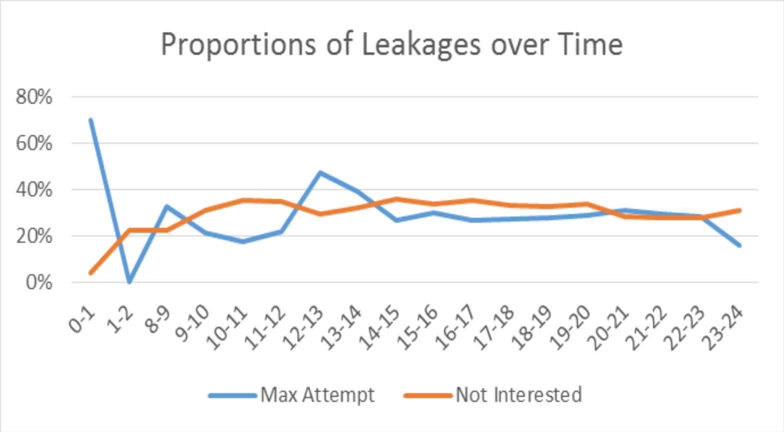
Feedbacks were recorded in two major categories for **Customer Driven Leakages**:

1. Call back, leakage study hour over hour
2. Not Interested, leakage study hour over hour



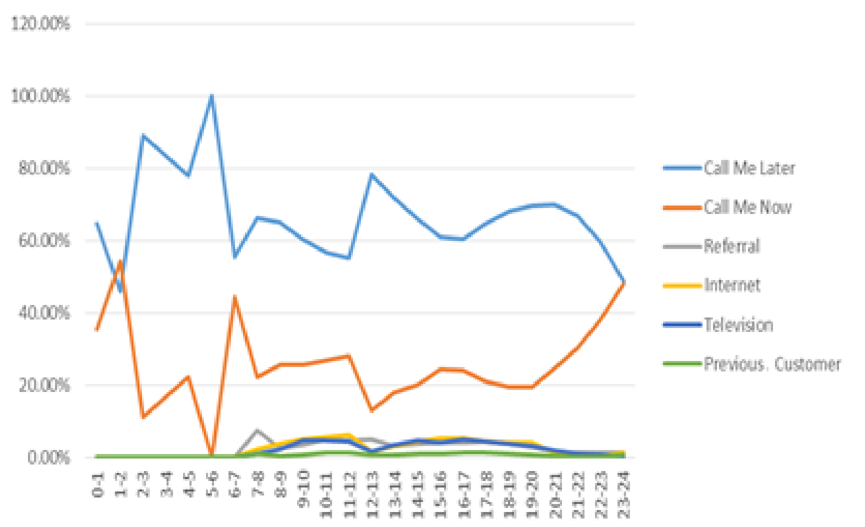
Data collected over more than 4 years showed:

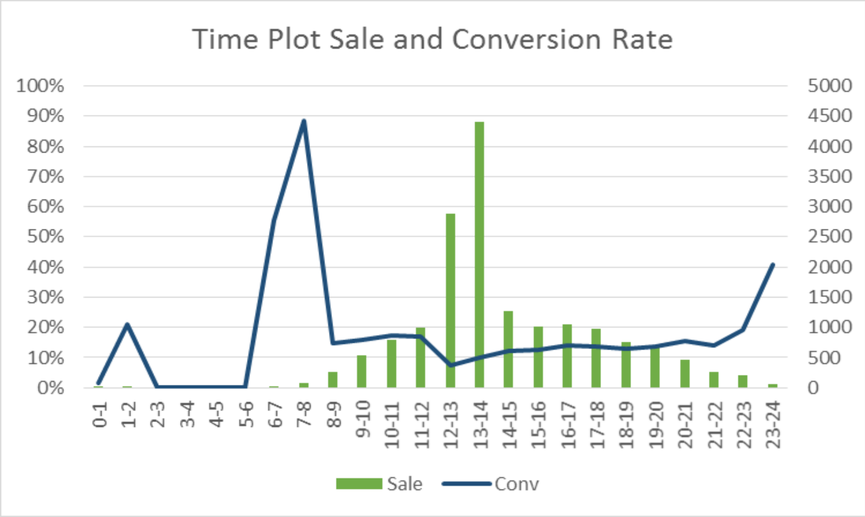
- Gradual declining behavior in customer response for 'Not Interested' which can be explained as availability of customers during particular hours ()
- Max leads are observed from 12:00–1:00 PM followed by 1:00–2:00 PM and 2:00–3:00 PM



Customer Driven Leakage: Call Timing Analysis

- Month of **October** witnessed maximum sale. In **December** sale sinks to lowest
- Maximum ‘Call me later’ request – between 12:00 - 1:00 PM
- Maximum ‘Call me now’ – early morning 6:00-7:00AM and late evening post 7:00PM
- Conversion rate has been observed increasing pattern after 12:00 PM.
- Initial peak of conversion rate between 7:00-8:00AM am was due to overnight corrections.





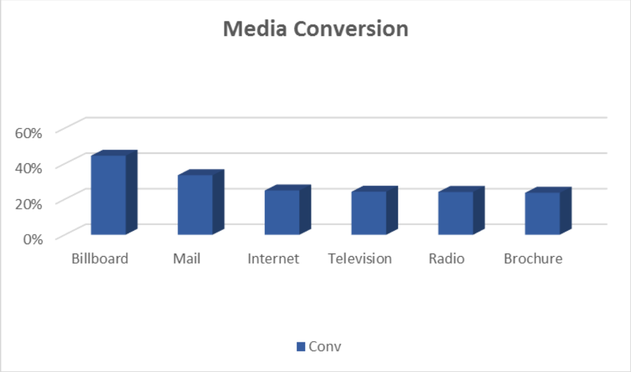
Customer Driven Leakage – Dispositions Analysis

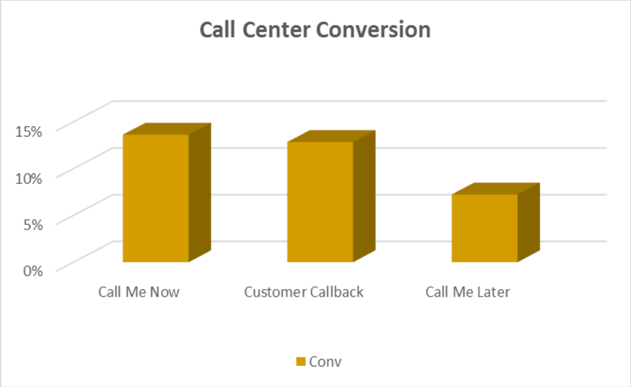
- 50% of leads and 37% of the customers (who finally bought the policy) found the products costlier than market.
- Only 8% Indecisive leads (shopping around/need more information/not ready for decision at first call) resulted in “Not interested”. This also meant that we were doing good in helping customers make decison
- 36.86% of the leads did not provide any initial comment (i.e. at first connect). Silver lining is 68% of customers (who finally bought the policy) did not provide initial comments either!
- Strong correlation between customers who expressed interest vs. acted upon i.e. prospects who expressed interest for more than product actually bought more than one product!

Product Affinity -Bundled Product		
No. of Product	Preference Shown	Finally Bought
2	82%	76%
3	15%	20%
4	3%	4%

Opportunity Leakage – Call Origin Analysis

- Company Affinity/ loyalty had best conversion rate of 31% followed by Media 25%
- Higher focus on Customer Service and Satisfaction was required as 15.14% of total policy sold comes from Prior and Existing Customers and Referrals. Conversion rate was also significantly higher presenting good opportunity to sell renewal contracts to prior customers and Cross selling to current customers
- Conversion rates of call origin as Mail, Billboard & Brochure are on higher side but total leads generated are less, so we have scope of generating more revenue by increasing lead flow
- Conversion rates of 'call me later' is very low but lead flow is on higher side, but its contribution to total policies sold is 44.3%: we can appoint a specialized team which is efficient in converting call me later customers





Key Findings – Stage 2

- Sentiment analysis categorically inferred that customers are expecting more than what they were getting. Customer service, price points and experience during the sale call being most important
- Strong correlation between customers who expressed interest vs. acted upon.
- Max sale in Jan, Mar & Oct, least in May & Dec
- Most favorable time to call b/w 11:00AM-12 noon & 2:00PM–4:00PM, Least favorable time 6:00-8:00PM, Least conversion rate b/w 12:00 Noon-01:00 PM
- Company’s loyal/associated leads had best conversion rate (32%) followed by Media 26%
- Only 8% Indecisive leads (shopping around/need more information/time) ended up “Not interested
- 35% of the leads did not comment at first connect. But, 63% of them finally bought the policy
- Agents were not able to use Pre qualifier questions or best product bundle (except ‘Accident-Cancer’)
- Agents lacked in identifying ‘Appetite’ vs. ‘Non Appetite’ risk exposures
- Leads from the most spent avenues were lesser than expected

Opportunities & Recommendations

Leading edge technology. Innovation driven.

Opportunities and Recommendations

1. Sales could be increased through implementation of one or a combination of following strategies:
   - Improving efficiency of certain agents and/or aligning right expertise on the right call
   - Picking right time for the call
   - Offering right product basket
   - Increase potential customer base by minimizing leakages
   - Rationalization of Marketing Spends

2. Devising effective training plan to bridge the significant performance gap among the agents
3. Benchmarking Prices to address the concerns of 53% leads. If price change not advisable, act on perception correction
4. Capture demographic data on ‘Pre-qualification Stage’. Companies lose very critical information!!!
5. Perform SWOT on South (strong) region and utilize the insights on other regions

Opportunities and Recommendations

6. Optimize call output by careful selection of time slots and incentives. For example,
   - If possible, plan ‘Offers’ for the month of May/June and December (low sale months)
   - Avoid ‘Max Attempt’ prone call b/w 6–8 PM, can be planned between 11–12 AM and 2-4 PM
   - Explore the possibility of offering “Happy Hours” or discounts (between 1;00-2:00 PM

7. Increase ARPU/CLTV by offering Product basket, e.g.
   - First choice ‘Accident’ - offer ‘Cancer’ and ‘Juvenile life’ to female prospect
   - First choice ‘Cancer’ - offer ‘SHE’ to male as well as female prospects.
   - First choice ‘Life’- offer Accident (followed by ‘Cancer’) to males (~60%) & females (~40%)
   - First choice ‘SHE’ - offer ‘Accident’ to male (60%) & female (40%)
   - If a female prospect already has ‘Accident’ cover, SHE+Cancer could be a good bundle
   - If the child/Juvenile Life is a male, Offering Juvenile life may have 70% better success rate
   - If prospect is a 40+ female, offering ‘Cancer’ product may have twice better conversion chances than male

Opportunities and Recommendation

8. Marketing Campaigns/Call origin:

– Continue focusing on Company associated/loyal customer. Start focusing on leads from 'Media'

– Start campaign to generate leads through Mail, Billboard & Brochure a potential relatively untapped

– Assign 'Call me later' cases to experienced callers as despite low lead inflows, but its contribution to total policies sold is 52%

– Place higher focus on customer service and satisfaction as 14.5% of total policy sold comes from Prior and Existing Customers and Referrals. Conversion rate is also significantly higher.

Observation, Suggestions and Questions…

SUTHERLAND
GLOBAL SERVICES®

SUTHERLAND
GLOBAL FOUNDATION