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		

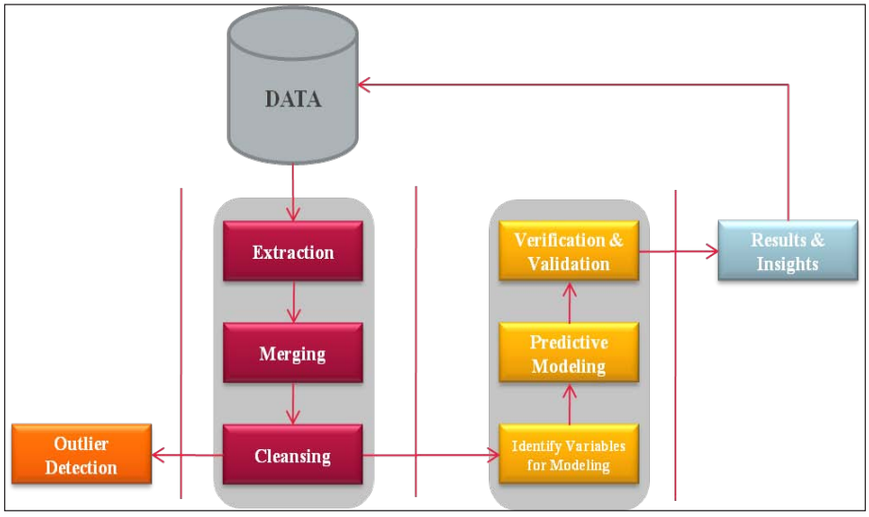

Figure 1: Data Analytics Flow

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

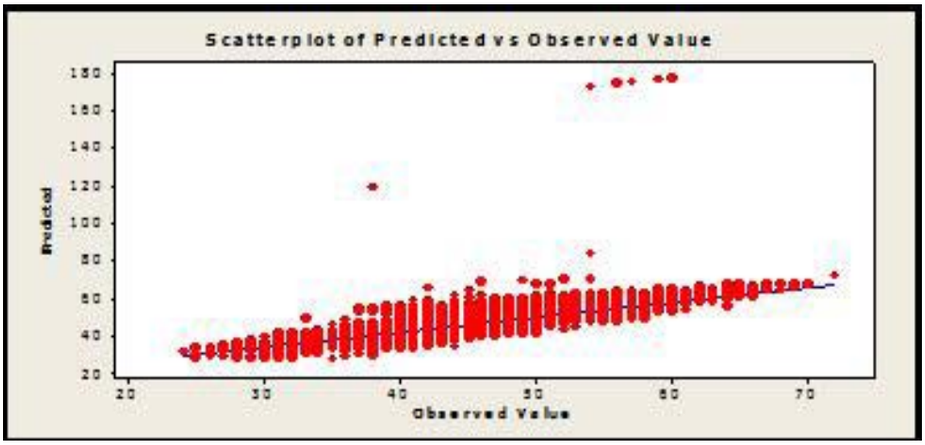

Figure 2. Scatter plot of Predicted vs. Observed value

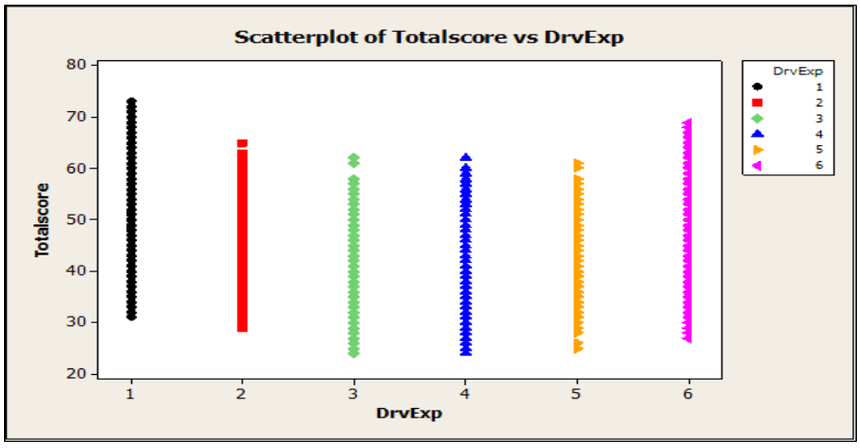

Figure 3.a. Total Score vs. Driver Experience

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.

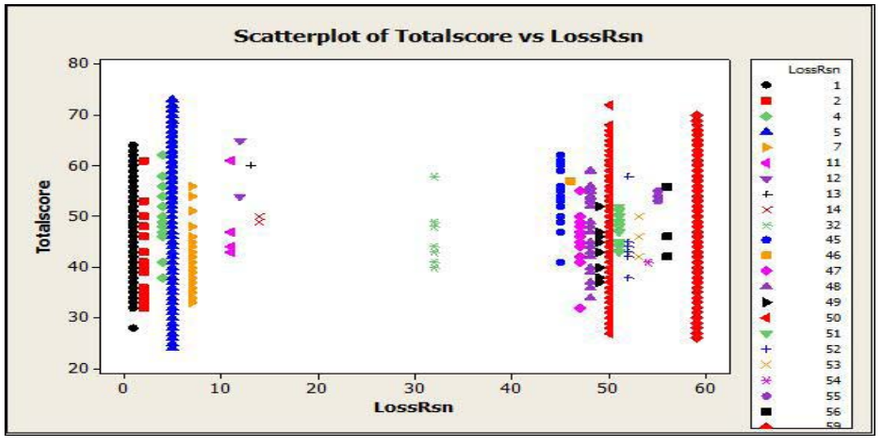

Figure 3.b. Total Score vs. NOL

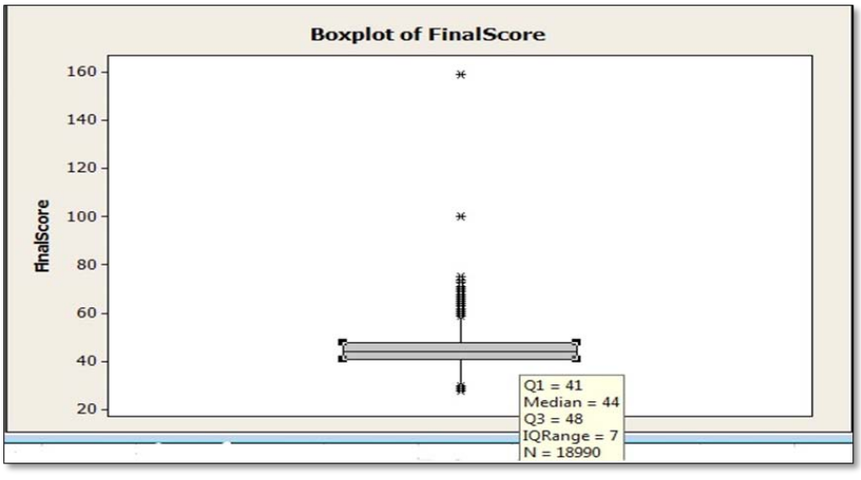

Figure 4. Box-plot of Final Score

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.