A Feasibility Study on Automating the Automotive Insurance Claims Processing

Jayanta Basak
IBM India Pvt Ltd
India Research Lab
bjayanta@in.ibm.com

Desmond Lim
IBM Singapore Pte Ltd
IBM Software Group
limd2@sg.ibm.com

Abstract

In this paper, we present a feasibility study in automating the insurance claims processing system related to automotive sector. In automobile insurance claims processing, usually an adjuster is sent to specific location from where the claim has been registered to physically inspect the validity of the claim. Sending adjusters for each and every such claim imposes a cost factor. With the increased number of such claims it often becomes very difficult to physically inspect each claim with a limited number of adjusters. Here we categorize certain claims where physical inspection is unnecessary and these claims can be processed without sending any adjuster (we call these claims as 'fast-track' claims). The challenge of labeling any claim as a 'fast-track' claim is that in the past historical data no claim has been identified as 'fast-track' claim. The domain experts have certain knowledge about 'fast-track'ness of a claim, however, it is very difficult to articulate that knowledge in the form of crisp rules. Rather if certain rules are articulated then the domain experts can validate whether such rules agree with their expertise or not. Therefore, our objective is to label a subset of claims as 'fast-track' claims and extract the rules for labeling any new claim as 'fast-track' or not such that the rules must agree with the domain knowledge of the experts. We consider the delay in claims processing as an indicator variable, and learn decision tree structures for various different thresholds on the delay. We arrive the final decision tree structure where the total cost of false positive and false negative samples is minimized on the labeled data. Finally we validate the rules extracted by the decision tree with domain experts and observe that the decision tree brings out the domain expertise as the top set of rules.

1 Introduction

In this paper, we present a feasibility study on claims reengineering process performed for an Indian automobile insurance company. At present in the specific claims processing system, whenever a claim is made, the relevant descriptions are entered into the system and then an adjuster is sent to verify the claims. Currently, all claims are subject to inspection by claim adjusters and the amount of indemnity payment for a claim is determined as part of the adjustment process. It is commonly believed that a claim adjuster can reliably process around six claims per day and no more than twelve without there being a decline in the quality of the inspections process. The objective of this specific feasibility study is to automatically determine the need of sending an adjuster into the specific claimant location. Our aim is to build a model to predict if the insurance company should send an adjuster to decide the indemnity amount or the company can decide the amount from the claims description entered into the system. The main bottleneck in this feasibility study is the absence of labeled information in the past historical data in the database. The database stores the indemnity amount paid to the claimants in the past after the adjustment process since indemnity amount was never decided directly from the claims description without sending an adjuster.

Not sending an adjuster saves certain cost for the company whereas it may incur certain business risk of exaggerated claims or fraudulent claims. Our objective is to detect such fraudulent claims with a constraint that the past historical data does not provide us any direct labeled samples about which were fraudulent claims in the past. In the literature several investigations have been reported on statistical and data-mining based techniques for fraud detection [4, 8, 18, 9, 13, 3, 12]. These studies include several techniques for fraud detection in unsupervised manner, however, these kinds of fraud mostly include the analysis of certain behavioral pattern (such as computer intrusion detection, credit card fraud) and detecting

15th International Conference on Management of Data COMAD 2009, Mysore, India, December 9–12, 2009

the anomalies in those patterns. In our context of automobile insurance sector, any exaggerated claim can be considered as a fraud. The exaggerated claims can happen in many different ways and not just for any bad intention of the specific claimant. The claim registration can be made from a car garage and that specific agent can also be responsible for exaggerated claims. In other words, the nature of fraud in our specific case study is different from those addressed in the general literature of fraud detection. In the context of automobile insurance sector, there exist several investigations on fraud detection [11, 19, 7, 17, 6, 16, 1] which include investigations using various predictive modeling techniques to label certain claim as fraudulent claims. These investigations address various different aspects of automobile insurance claims fraud such as exaggerated claims (the specific concern of our paper), bodily injury, car theft. However, in almost all these approaches, a basic assumption is that the past historical data contains the information about false or fraudulent claims. Therefore, most of the predictive modeling techniques used in these investigations pertain to supervised modeling or associating certain supervised information to observed groups of claims.

In our feasibility study, we do not have any specific supervised information about a claim being false or fradulent. This is due to the fact that the past historical data stores only the adjusted amount paid to the claimant. In order to label any sample in the historical data as fraudulent claim, each record needs to be manually checked to verify whether the estimated amount from the original claim record matches with the adjusted amount or not. It was not feasible to manually annotate the huge volume of records by the domain experts, and our objective is to obtain an automated mechanism to label the past records. We call the claims which are not exaggerated as 'fast-track' claims since in such cases, it is unnecessary to send an adjuster to physically inspect the claims. In the absence of any labeled samples, the task involves completely unsupervised learning or estimation. However, even if we cluster the claims it is not straight-forward to analyze which cluster actually corresponds to 'fast-tracked' claims.

The domain experts have certain knowledge about the nature of the claims which can be processed directly, however, it is very difficult to articulate that domain knowledge into hard-coded rules. Therefore, it is very difficult to assign any cluster as "fast-track" or not-fast-track (exaggerated claim) according to the domain expertise. The second problem in this domain is the validation. Since there is no labeled data, it is not possible to estimate the accuracy of the system with respect to making a decision. The only way of validating the system is to obtain the certification of the domain experts, i.e., whether the set of claims labeled as 'fast-track' by the system actually makes any business sense or not according to their domain knowledge. Again the domain experts have certain knowledge which is not actually expressible as any information regarding clusters or cluster centers.

In this paper, we use a decision tree [14, 5] to estimate the likeliness of claims being fast-tracked. We choose a decision tree learner mainly to validate the rules extracted by the decision tree by the domain experts. It is possible to construct unsupervised decision trees [2, 15, 10] on the claims data. However, the rules extracted by the unsupervised decision tree only reveal certain cluster boundaries and do not directly correspond to the domain expertise regarding any claim being exaggerated.

In the past historical data, there are two key indicators (dependent variables) which we use in building a decision tree. These two variables are indemnity amount paid after adjusting the claim and the delay (processing time) in processing the claim. A basic assumption is that a straight-forward claim which can be processed without the need of any physical inspection should be simple enough to require less processing time and the indemnity amount should not be very high. With the knowledge of the claim processing industry, we fix certain threshold on the paid amount (which is also in accordance with the existing regulations). Therefore any claim with an amount greater than the threshold has to be physically inspected. Moreover, any claim involving bodily injury must go through physical inspection according to the regulation. We consider a large number of independent variables reflecting the demographics, nature of the claim, type of automobiles etc. We first perform data cleansing and then define certain derived variables using the available data. We then construct a decision tree with the dependent variable of claim processing delay as an indicator variable. The basic concept is that we fix a threshold on the claim processing delay such that any claim in the historical data with a delay less than the threshold can be considered as a case of 'fast-tracking'. However, fixing the threshold on the delay is a non-trivial task, and we use a balanced decision tree (in terms of false positive and false negative) to set this threshold. Once we obtain the decision tree with rules comprising of the independent variables, we validate the tree with the domain experts. We observed that the rules extracted by the tree not only perfectly matches with the domain expertise but also represents the domain expertise in a much more quantitative manner.

2 Data Cleansing and Claim Attribute Selection

Merging entries: In the database, usually one claim has more than one entry. A claim is registered whenever the claim is made with a claim identifier along with the date (claim date or loss-reporting-date) and time and other details such as vehicle make, location, policy identifier, claim type, claim description. At every processing stage including the physical inspection and part payments, separate entries are made into the database with the same claim identifier. Finally, a final payment date along with the last payment is entered for a claim when it is settled. We merged all the entries for a specific claim with the same identifier, and considered the date of first entry as the claim date and the final payment date as the settlement date. We added all part payments and the final payment as the total indemnity paid to the claimant. All other entries such as vehicle make, model, location, policy identifier, claim description etc. remains the same as the first entry.

Invalid Claims: We considered the policy database for detecting any anomaly or false claim registered. If a claim date happens to be before the start date of the corresponding policy then the claim is invalid and we removed those entries from the claim database. Similarly, if a claim date (loss-reporting-date) happens to be after the end date of the corresponding policy then also the claim is invalid and we removed such claims as invalid claims.

Derived Attributes: The claims database and policy database after processing contains various independent attributes such as claim dates, cause of claim, claim type (whether bodily injury or vehicle damage or both or vehicle theft etc.), vehicle makes, vehicle models, location, policy holder's demographics, loss reporting date. We do not consider direct entries of certain attributes. For example, we do not consider the loss-reporting-date and policy-date directly. We define certain attributes which we consider as the independent variables in our modeling. In order to consider the policy dates and the loss-reporting-date together, we consider the difference between the loss-reporting-date and the policy start date as one variable, and the difference between the loss-reporting-date and the policy-end-date as another independent variable. We define these variables mainly because in the automotive sector, certain exaggerated claims happen near the policy end date. Similarly certain past accidents are claimed in roll-over policies i.e., at the very beginning of a policy certain exaggerated claims may happen. We also define a binary variable to indicate if the loss location is close to the settling office. In many cases, domain experts mentioned that fake claims have a tendency of reporting a loss location far apart from the settling office so that physical inspection can be bypassed. Often in the actual claims, the policy holder is not the same person as the claimant i.e., claimant name differs from the policy holder's name. We define a binary variable to indicate if the claimant name is the same as the policy holder's name. In the set of independent variables we do not consider certain attributes for modeling. These include several variables such as the actual claim identifier, claimant's name, policy holder's name. The policy database stores the vehicle make (manufacturer) and model names. However, often a model has several variations. For example the make and model can have more than one categories (e.g., includes power steering or not, power window or not etc.). We introduced two other variables to reflect the mean price and the variance of the price for each make and model. We also normalized the prices with price depreciation factor to define the mean price and price variance. The make and model names are in the form of unstructured text in the database and we transformed them to structured text for further modeling. Finally, we use a list of 30 independent variables for each claim including

- claimType: if the claim is for bodily injury/death/property damage/theft etc.
- coverage: Total policy coverage
- polIssueOffice: Office from where insurance policy has been issued
- lossLocation: Location of loss
- settlingOffice: Location of office where claim will be settled
- causeLossText: Cause of loss
- VehMakeName: Name of vehicle make
- VehMdlName: Name of vehicle model
- claimantNamePolicyHolder: Binary variable indicating if Claimant name is same as Policy Holder's name
- lossLocSettlingoff: Binary variable indicating if loss location is same as settling office
- lossDateLossReported: Difference between Loss date and loss reported date
- lossReportedPolicyStartDate: Difference between Loss reported date and Policy start date
- lossReportedPolicyEndDate: Difference between loss reported date and Policy end date.
- MeanPrice: Average Price of vehicle model
- PriceVariance: Variance of the price

We considered two indicative dependent variables namely,

- dateLossReportedDate: Difference between date of indemnity payment and date of loss reported
- indemnityPaid: Total amount paid to the claimant after settlement

The first dependent variable indicates the delay in processing a particular claim and the second variable indicates the severity of the claim.

3 Decision Tree Modeling

Our main objective is to model the data such that certain claims can be labeled as 'fast-track' and the models can be described in terms of certain rules which agree with the domain expertise. As we mentioned before, it is very difficult to construct clustering based approach on the independent variables because in that case we have to associate the delay with each cluster and there is no obvious reason for choosing a cluster as 'fast-track'. Moreover, the clusters have to be expressible in the form of certain crisp rules such that they can be validated by the domain experts. It is possible to construct unsupervised decision trees [2, 15, 10] directly on the independent variables. However, unsupervised decision trees reveal certain rules describing the cluster boundaries which do not directly translate into the insurance domain expertise.

As a first step, we label all the claims involving bodily injury and car theft as 'not-fast-track' claims according to the regulations of the insurance company [1]. Since these two conditions gives us obvious rules (independent of any other variable), we eliminate all the claims involving bodily injury or car theft.

As a next step we model to obtain two most suitable thresholds, one on the amount paid to the claimant and the other on the delay in processing a claim such that any claim having an amount greater than the first threshold needs to be inspected by an adjuster, and any claim with a processing time greater than the second threshold is a candidate for physical inspection. The insurance company has a specific regulation that if the claim amount goes beyond a specific threshold [2] then the specific instance must be physically inspected. Therefore, we label all the existing claims in the database with an 'indemnityPaid' greater than the specific threshold (provided by the insurance company) as 'not-fast-track' claims.

In the third step, we find the most suitable threshold on the delay (processing time) in claim settlement [3]. There was no obvious guideline from the insurance company about any specific threshold on the delay, and we obtain a most suitable threshold on the delay in our model. The objective of finding out a threshold on the processing time is that any claim with a delay less than the threshold can be considered to be simple enough such that these claims can be considered as 'fast-track' claims. If we are able to derive such a threshold then we can label the samples and subsequently build a decision tree to generate the rules. As we plot the distribution of the delay of all claims in the database we observe that the distribution is completely unimodal in nature (mostly Poisson distribution) and there is no obvious choice of any threshold as in the case of a bimodal distribution [4].

In order to select a threshold, first we consider an arbitrary threshold T. With this threshold, we label the samples such that the samples with a delay less than T are positive ('fast-track') and the others are negative ('not-fast-track'). We then build a decision tree on these labeled samples. The decision tree labels certain samples as false positive and certain samples as false negative on the data labeled using the threshold T where a positive indicates a 'fast-track'. The false negative samples indicate a cost of sending an inspector while any physical inspection is unnecessary. The false positive samples indicate a cost of exaggerated claim because of not sending an adjuster. Let c_1 and c_2 be the costs of false positive and false negative respectively (can be determined by the actual business unit) then we should choose a threshold such that $c_1FP + c_2FN$ is minimum where FP and FN denotes the number of false positive and false negative samples. It was very tricky to decide any ratio of c_1 and c_2. These two costs are not comparable because the first involves business risk and an unsafe error although it saves money for the business unit. On the other hand the second one is a safe error for the business unit (in terms of risk) but incurring additional cost to the business unit. In the absence of any quantitative comparable measure, we considered both of them equal (both of them are equally undesirable).

Therefore, we select a threshold such that $FP+FN$ becomes minimum. Interestingly, we observe that as we increase the actual number of positive samples by increasing the threshold the number of false positives increases drastically. Similarly, as we increase the number of negative samples by lowering the threshold, the number of false negatives increases in the decision tree model. In the extreme condition when almost all samples become positive or negative, the total number $FP + FN$ reduces (due to high biased situation). Thus we obtain a region where we observe a convex nature of the $FP + FN$, and we choose the threshold for which $FP+FN$ is minimum. Interestingly, we also observe that under such condition, the values of FP and FN becomes almost equal.

4 Results

We had a database with more than 100000 entries with multiple entries for each claim. After data cleansing (including removal of invalid entries and merging multiple entries for each claim as described in Section 2), we remove all the claims involving bodily injuries and

[1] In the case of bodily injury and car theft, there has to be police inspection and insurance company has to physically inspect the incidence

[2] The actual amount cannot be revealed due to confidentiality

[3] It is possible to judge if a claim in the past was exaggerated or not by comparing the claimed amount and the indemnity paid to the claimant. However, the claimed amount of the past claims are not stored in the database because all claims in the past had been physically inspected, and only the amounts settled by the adjusters were entered into the database.

[4] The exact histogram cannot be provided due to restrictions

car theft. Finally, we have approximately 35000 entries with 30 independent attributes (including the derived variables). The dataset contains all claims from various different regions across the country.

By labeling the samples (as described in Section 3), we obtain a 62% accuracy of the decision tree model on all labeled samples. As we construct decision trees for every region separately, the decision trees collectively provide an accuracy of approximately 80% on the labeled samples. However, these accuracies only reflect how far the decision trees are able to agree with the chosen threshold which may not reflect the actual accuracy after deployment of the system. We therefore validated the rules obtained from the decision trees with actual domain experts. We observed that the final decision trees bring out the knowledge of the domain experts as the top rules irrespective of the regions. We provide four such findings as follows:

1. - **Hypothesis:** Claims made of rollover policies early in their lifetime are more likely to be exaggerated.
   - *Finding (Decision Tree):* If the gap between loss-reporting-date and policy-start-date is less than certain threshold then it is always flagged as ‘not-fast-track’, the threshold is decided automatically by the decision tree.
2. - **Hypothesis:** Claims made in a city geographically distant from the actual loss location are likely to be exaggerated.
   - *Finding (Decision Tree):* If the loss-location is not closest to the settling office, then it follows a path in the tree that is more likely to be ‘not-fast-track’.
3. - **Hypothesis:** Claims not made by the policy holder, more specifically by non-approved garages, are more likely to be exaggerated.
   - *Finding (Decision Tree):* If the claimant name is not same as the policy holder’s name then the claim is most likely to be ‘not-fast-track’.
4. - **Hypothesis:** Some descriptions of reported damage are more likely to be exaggerated than others
   - *Finding (Decision Tree):* ‘causeLossText’ which is a structured text in the claims database plays an important role in making a decision about ‘fast-track’ or ‘not-fast-track’. We obtain certain reasons provided by the ‘causeLossText’ as the top rules which completely agree with the actual domain expertise.

There are several other rules that the decision trees extract concerning the region, vehicle make and models, and demographics specific to regions which perfectly matches with the domain expertise; however, we do not reveal those rules here due to certain constraints.

5 Conclusions

In this paper, we presented a feasibility study that we performed in automating the automobile claims insurance processing system for an Indian company. In this feasibility study, we modeled the claims in the automotive insurance sector such that certain claims can be processed without involving any physical inspection. In the absence of any labeled samples in the past historical data (every claim has been physically inspected), our task boiled down to finding out the suitable threshold on the delay in claims processing such that any claim processed with a delay less than the threshold can be considered as a simple straight-forward claim and automatically can be processed without physical inspection. Since the distribution of the delay of all claims is completely unimodal in nature (similar to Poisson distribution) there was no obvious choice of selecting any specific threshold. We learned decision tree structures for various different thresholds such that for each threshold we derive a labeled sample set and learn one decision tree. We then obtain the decision tree with the least number of errors committed (in terms of false positive and false negative) and also observe that the specific decision tree is balanced in terms of false positive and false negative. We then validated the rules extracted by the decision tree by the actual domain experts of the insurance company.

We consider two different costs in modeling, one for the false positive and the other for the false negative. These two costs are not comparable in the sense that the first involves risk and an unsafe error although it saves money for the business unit. On the other hand the second one is a safe error for the business unit (in terms of risk) but incurring additional cost to the business unit. In the absence of any comparable measure, we considered both of them equal. However, the rules can further be refined with certain preference on the cost parameters which can be used as user-defined parameters and provides certain additional utility to the domain experts. However, the performance of the system can be improved if the actual labeled samples (physically inspected claims with actual labeling of ‘fast-track’ or ‘not-fast-track’) are available from the domain experts which pertains to the predictive modeling in supervised learning.

References

[1] M. Artís, M. Ayuso, and M. Guillén. Detection of automobile insurance fraud with discrete choice

models and misclassified claims. *Journal of Risk and Insurance*, 69:325–340, 2002.

[2] H. Blockeel, L. De Raedt, and J. Ramon. Top-down induction of clustering trees. In *Proc Int. Conf Machine Learning (ICML 1998)*, pages 55–63, 1998.

[3] R. J. Bolton and D. J. Hand. Unsupervised profiling methods for fraud detection. In *Proc. Credit Scoring and Credit Control VII*, pages 5–7, 2001.

[4] R. J. Bolton and D. J. Hand. Statistical fraud detection: A review. *Statistical Science*, 17:235–255, 2002.

[5] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Chapman & Hall, New York, 1983.

[6] P. L. Brockett, X. Xia, and R. A. Derrig. Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. *Journal of Risk and Insurance*, 65:245–274, 1998.

[7] R. A. Derrig. Insurance fraud. *Journal of Risk and Insurance*, 69:271–287, 2002.

[8] T. Fawcett and F. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1:291–316, 1997.

[9] Y. Kou, C.-T. Lu, S. Sirwongwattana, and Y.-P. Huang. Survey of fraud detection techniques. In *Proc. IEEE Intl. Conf. Networking, Sensing & Control; Taipei, Taiwan*, pages 749–754, 2004.

[10] B. Liu, Y. Xia, and P. Yu. Clustering through decision tree construction. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, pages 20–29, McLean, VA, USA, 2000. ACM.

[11] J. M. Pérez, J. Muguerza, O. Arbelaitz, I. Gurrutxaga, and J. I. Martn. Consolidated tree classifier learning in a car insurance fraud detection domain with class imbalance. In S. S. et al., editor, *Lecture Notes in Computer Science: Pattern Recognition and Data Mining; Proc. Intl. Conference on Advances in Pattern Recognition, Bath*, volume 3686, pages 381–389. Springer, Berlin/Heidelberg, 2005.

[12] C. Phua, D. Alahakoon, and V. Lee. Minority report in fraud detection: classification of skewed data. *SIGKDD Explorations*, 6:50–59, 2004.

[13] C. Phua, V. Lee, K. Smith, and R. Gayler. A comprehensive survey of data mining-based fraud detection research. Technical report, Clayton School of Information Technology, Monash University, Victoria, Australia, 2005.

[14] J. R. Quinlan. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence*, 4:77–90, 1996.

[15] J. Struyf and S. Dzeroski. Clustering trees with instance level constraints. In *Proc European Conf Machine Learning (ECML 2007)*, pages 359–370, 2007.

[16] S. Viaene, G. Dedene, and R. Derrig. Auto claim fraud detection using Bayesian learning neural networks. *Expert Systems with Applications*, 29:653–666, 2005.

[17] S. Viaene, R. Derrig, B. Baesens, and G. Dedene. A comparison of state-of-the-art classification techniques for expert automobile insurance fraud detection. *Journal of Risk and Insurance*, 69:373–421, 2002.

[18] R. Wheeler and S. Aitken. Multiple algorithms for fraud detection. *Knowledge-Based Systems*, 13:93–99, 2000.

[19] J.-H. Yoo, B.-H. Kang, and J.-U. Choi. A hybrid approach to auto-insurance claim processing system. In *Proc IEEE Intl. Conf. Systems, Man, and Cybernetics: 'Humans, Information and Technology'*, volume 1, pages 537–542, 1994.