SMARTR: A Framework for Early Detection using Survival Analysis of Longitudinal Texts

Jean-Thomas Baillargeon and **Luc Lamontagne**
{jean-thomas.baillargeon, luc.lamontagne}@ift.ulaval.ca
University Laval, Québec, Canada

Abstract

This paper presents an innovative approach to the early detection of expensive insurance claims by leveraging survival analysis concepts within a deep learning framework exploiting textual information from claims notes. Our proposed SMARTR model addresses limitations of state-of-the-art models, such as handling data-label mismatches and non-uniform data frequency, to enhance a posteriori classification and early detection. Our results suggest that incorporating temporal dynamics and empty period representation improves model performance, highlighting the importance of considering time in insurance claim analysis. The approach appears promising for application to other insurance datasets.

1 Introduction

Most claims from the car insurance industry are straightforward to settle. The damage to the car's body generates benefits that are easy to predict. These prevalent claims are part of the loss an insurer can foresee from year to year in the portfolio of its policyholders. Catastrophic claims, on the other hand, occur at unexpected moments, are of a completely different magnitude, and pose a danger to the company's financial health.

These costly claims result from the bodily injuries a policyholder will suffer during a car accident. These injuries can, in the most extreme cases, cause permanent damage to the policyholder, such as disability or amputation. In addition to ranging from $100,000 to several million dollars, their handling can span over many years, during which various experts try to agree on the settlement.

Early detection of such claims is desirable: although the original injuries have occurred, taking care of that policyholder can prevent risk deterioration that causes more significant costs. Furthermore, since these payments span several financial years, the actuaries need to adequately provision for future benefits so the money is reserved for the insured, not paid to shareholders as profits.

Our application attempts to detect expensive claims early in a privately held longitudinal textual corpus from a Canadian insurer. This corpus contains claim files comprising textual documents monitoring a claim's settlement process over time, which we believe is helpful in detecting expensive claims early.

The main contribution of this paper is the SMARTR model, an early classification model that uses a survival analysis model calibrated on text data. In our proposed model, adding a temporal aggregating layer and monthly padding improves the early detection time by, on average, 4 % without decreasing its classification performance.

This paper is divided as follows. We present related work for survival analysis and fields interested in early detection in Section 2. We then present the groundwork to include our dataset and survival analysis into a classification task in Section 3. Finally, we present the evaluation scheme, our models, and results analysis in Section 4.

2 Related work

Survival analysis aims to relate factors causing an event to the waiting time until its occurrence. Classic examples of using this analysis include evaluating the waiting time until a mechanical part fails or until a person dies. In the present paper, we model the waiting time between the occurrence of an accident and the moment it is identified as expensive.

Using a specialized neural encoder to generate representations used to calibrate a survival model is a familiar idea. A review of classical models was conducted by Baesens et al. (2005) for credit scoring. These models were set aside until the mid-2010s, when neural networks benefited from significant advancements. A more recent review presents

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 36–41

advances in machine learning survival models in Wang et al. (2019).

The first neural implementation of a Cox Proportional Hazard survival model (CPH) was presented by Faraggi and Simon (1995). In their work, the authors developed a network that offered automatic encoding of attribute interactions (Xiang et al., 2000). The next iteration of the neural CPH model, DeepSurv (Katzman et al., 2018), combines several architecture and methodology improvements. The better performance of this model demonstrates the ability of neural encoders to exploit complex interactions to calibrate a survival function.

However, these approaches and models exploiting survival functions are not designed to handle textual data and have yet to be evaluated with longitudinal textual data in the application of a costly claim identification problem.

The fraud detection field is interested in early detection (Liu et al., 2020; Xiao et al., 2023), but our problem differs from theirs as we have a gold label to trust and leverage to train a classification model. Medicine is also interested in early detection (Pan et al., 2020; Sungheetha et al., 2021); but the clinic uses cases that are seldom evaluated using time-varying covariates from a longitudinal study as our problem is.

Alternative approaches for early classification include adversarial training (Chapfuwa et al., 2020), where a loss function is calibrated to optimize the tradeoff between timeliness and performance. Although attractive, we prefer an approach that provides a risk evaluation framework that actuaries can leverage in insurance operations and processes.

Our model is inspired by the SAFE model presented by Zheng et al., which lacks the capacity to handle text data and is bound to inputs and labels produced at the same frequency (e.g. daily or monthly), two limiting factors to address our use case.

3 Methodology

This section describes the dataset used and the approach to classifying observations using a survival probability.

3.1 Dataset Used

The dataset used in our study contains over 70,000 claim files from a Canadian car insurer. We labeled those claims as expensive whenever the total payout is above $ 50,000 or normal otherwise. This threshold overlaps two business classification thresholds (basic and expensive) that account for 7% of the dataset, making this task more complex to solve than trivially using textual artifacts from business processes.

We partitioned the dataset into three folds, which respectively hold 80 %, 10 %, and 10% of the complete corpus and are used for training, hyperparameter search, and results purposes. Furthermore, we verified that each partition contained roughly the same proportion of positively labeled examples. These examples contain a longitudinal observation that monitors the evolution of a claim through textual conversations between actors in the claims settlement process. These actors include, among others, claims adjusters, lawyers, and doctors. Each claim contains, on average, 75 notes made of 128 words. These notes have different information values: some concern critical elements of the claim, such as the accident description or the insured's injuries, while others are merely administrative artifacts, such as a mention of a clerk transfer. Furthermore, the distribution of notes over time is non-uniform, so there can be several months without any notes or more than half the notes occurring within a single month.

Another critical aspect of our dataset is the mismatch between the severity label, assessed using monthly aggregated benefit amounts, and claim notes, which can occur at any time (non-periodic) and are kept individually (not aggregated).

The particular characteristics of our dataset are rare and make replication of our experiments impossible on open datasets.

3.2 Classification using Survival Probability

Classification with a survival probability requires alterations to the classification model, so it generates risk factors that allow a survival rate to be calculated. Instead of assigning a class, we use this rate to rank each claim according to its inherent risk, as per the calculated model.

By comparing their survival probability, we infer whether a claim is more likely to become costly than the others. We calibrate a decision threshold using claims from the hyper-parameter partition. For each of those claims, we evaluate their survival probabilities $S_T(t), t \in 0, ..., t^i$ at each time step t and rank the claims according to their probability of becoming costly. For each time step t, we seek the threshold value that optimizes the separation between the two classes according to the F1

Score, and we classify claims that have survival probability below this threshold as expensive.

3.3 Calculating Survival Probability

We model the probability $S_T(t)$ of a claim to survive the event (i.e., not transitioned to state) of exceeding the costly threshold during at time t using the function :

$$S_T(t) \;=\; P(T > t), \tag{1}$$

where T is the random variable of the waiting time before the claim exceeds the costly threshold; the smaller this quantity is, the more likely the claim is to have exceeded the threshold at time t. We use a non-parametric model to calculate the probability $P(T > t)$:

$$S_T(t) = e^{-\sum_{x=0}^{t} \lambda_x} \tag{2}$$

Equation (2) uses the instantaneous failure rate λ_t defined as:

$$\lambda_t = P(t < T \leq t+1 | T > t) \tag{3}$$

These risk factors λ_t are produced by a neural network trained on a special loss function presented in this section.

3.3.1 Objective Function to Optimize

In a survival framework, we train the network to maximize the likelihood of each observation to survive (or not) at time $T = t^i$, defined as:

$$\mathbf{L}(\mathbf{x}^i, t^i, c^i) = P(T = t^i)^{c^i} \cdot P(T \geq t^i)^{1-c^i}, \tag{4}$$

where the variables x^i ,t^i, and c^i are defined as follows:

- $\mathbf{x}^i$: the accumulation of textual content of notes for claim i at time $t = t^i$ used as input to compute the probabilities.
- t^i: the moment when the benefits of claim i exceeded \$50,000 (or when this claim was no longer observed).
- c^i: an indicator variable if the claim became costly during the observation period.

The formulation of $\mathbf{L}$ from Equation (4) must be adjusted to optimize early detection and integrate our survival model hypothesis.

We assume the accident can be identified as expensive before the claim exceeds the expensive threshold at time $T = t^i$ whenever enough indicators are accumulated in the interval $[0, t^i]$. This assumption replaces a traditional one from the survival framework; the probability $P(T = t^i)$, that the claim i becomes costly exactly at time t^i, is updated with $P(T \leq t^i)$, the probability the claim becomes costly before t^i . This adjustment is reflected in the objective function $\mathbf{L}^*$ we use.

$$\begin{aligned} \mathbf{L}^* &= P(T \leq t^i)^{c^i} \cdot P(T \geq t^i)^{1-c^i} \\ &= (1 - S_T(t^i))^{c^i} \cdot S_T(t^i)^{1-c^i} \end{aligned} \tag{5}$$

As we assume a non-parametric model and use (2) to define the survival probabilities $S_T(t^i)$, we can derive the loss function backpropagated in the network from Equation (5).

$$\mathbf{L}^* = (1 - e^{-\sum_{t=1}^{t^i} \lambda_t})^{c^i} \cdot (e^{-\sum_{t=1}^{t^i} \lambda_t})^{(1-c^i)}$$

The likelihood function $\mathbf{L}^*$ is converted into its log-likelihood ℓ^i version.

$$\ell^i = \big(\sum_{t=1}^{t^i} \lambda_t\big) - c^i \cdot \ln\big(e^{\sum_{t=1}^{t^i} \lambda_t} - 1\big)$$

Finally, we backpropagate loss function $\mathcal{L}$, which combines losses ℓ^i for each of the N claim files in the training dataset defined by :

$$\mathcal{L} = \sum_{i=1}^{N} \left[\big(\sum_{t=1}^{t^i} \lambda_{t^i}\big) - c^i \ln\big(e^{\sum_{t=1}^{t^i} \lambda_{t^i} - 1}\big) \right]$$

Although many λ_t are calculated for this formulation, only one loss value based on the ground truth variables t^i and c^i is calculated and backpropagated for each training example.

3.3.2 Generating the λ_t

The values for λ_t are generated by a neural network trained to minimize the loss function $\mathcal{L}$. We train a recurrent cell to produce hidden states h_t from a claim encoder layer and convert them into λ_t using the function:

$$\lambda_t = \text{softplus}(w_\lambda h_t) = \ln(1 + \exp(w_\lambda h_t)),$$

where w_λ is the weight vector of a fully connected layer of the same dimension as the h_t, also learned during training.

4 Experiments

This section presents our evaluation scheme, models, baselines, and result analysis.

4.1 Evaluation

We evaluate the models on two axes: classification performance and detection speed. The basis of our evaluation is an iterative prediction of claim severity using an incrementing number of notes, mimicking a claim adjuster's work. We iteratively infer every claim class from the test dataset using the first 20 notes and then increment the number by steps of 20 up to 160 and 175, 200, 250, 300, 400, and 1000 (all notes) afterward.

We present two metrics for each model. The first one is the F1 Macro score (F1 Macro) of the a posteriori classification performance calculated using 1000 notes. This metric presents the model's capacity to exploit the complete longitudinal sequence information while addressing the light imbalance problem of our dataset. The second metric is the average proportion of notes the model requires to detect expensive claims correctly at the earliest time (ED). We compute this statistic by comparing the earliest time claims were correctly classified and the number of notes in the claim file when it reached the $50 000 threshold. We obtain the earliest time by iterating through all generated predictions (20,40,...,1000).

Both statistics are calculated by averaging results from ten runs and are presented with their 95% confidence interval when applicable.

4.2 Our Models

Our models leverage claims notes encoded by a RoBERTa transformer model (Liu et al., 2019), further pretrained on the Masked Language Modelling and Same File Prediction tasks as described in (Baillargeon and Lamontagne, 2024), and combine the resulting `[CLS]` tokens with LSTM cells. We propose and evaluate two models. The first is an adapted version of SAFE, and the second implements the capacities to handle the timing mismatch found in longitudinal data.

SMART The Survival with Maximum Aggregated Risk from Texts (SMART) model is the closest comparable to SAFE and is usable in our use case. As the latter cannot be used in our use case due to the time mismatch between notes and class label discussed in Section 3.1, we minimize the architectural impact to address this issue by using λ_t equal to the maximal risk factors generated for each note that belong to the same month.

SMARTR The Survival with Monthly Aggregated Risks from Texts Representations (SMARTR) model extends the SMART model with an additional layer that allows the construction of the time-varying covariate representation within the neural network. We present this architecture in Figure 1.

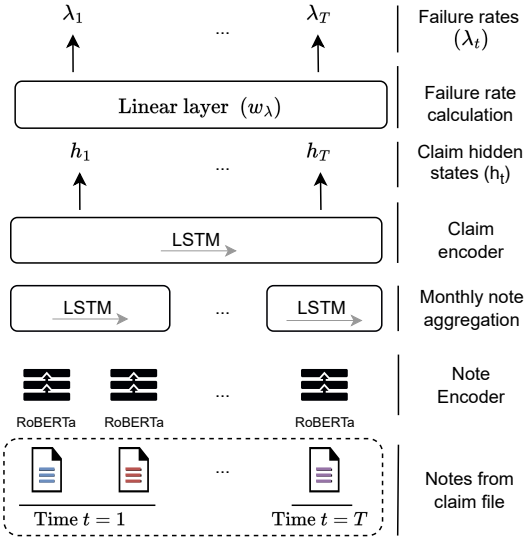

Figure 1: SMARTR model architecture

4.3 Baselines

To evaluate our model performance, we compare them to two baselines.

Logistic regression is a classic classifier that uses the Bag of Word representation of the claim to infer its class. In the early detection use case, this representation is generated using texts from up to the defined (20, 40, ..., 1000) reduced number of notes. This method is deterministic and does not generate confidence intervals on its results.

M-LSTM (Multi-source LSTM) is a neural classification model that uses an LSTM trained with the cross-entropy loss to capture time-varying covariates of the claim, presented in Yuan et al. (2017). In early detection use cases, hidden states at previously defined steps (20, 40, ..., 1000) are used for classification purposes.

This section presents results from our evaluation scheme for different cross-section analyses. The first evaluates the relevance of addressing the input and label mismatching issue found in our dataset by adding an embedding that represents passing time to months without any notes and of learning parameters to encode text inputs into a time-varying

covariate of a claim. This embedding is a zero-filled vector of 768 dimensions. The second one compares the results from the best configuration of SMART and SMARTR to the baseline models.

4.4 Results

4.4.1 Our Models Configuration Selection

We compare our model's performances using Table 1 results. This table presents the performance metrics presented in Section 4.3 for our two models.

Model	F1 Macro (%)	ED (%)
SMART*	77.25 ± 0.79	48.28 ± 1.50
SMART	79.51 ± 1.20	49.55 ± 0.62
SMARTR*	79.24 ± 0.51	**46.97** ± 0.98
SMARTR	**81.16** ± 0.25	**47.37** ± 0.86

Table 1: Classification Performances of Different Configurations of our Models, * indicates model not trained with the passing time embedding

By analyzing the confidence intervals overlap pattern, we conclude that adding a vector that models time passing improves a posteriori detection. However, ED results do not differ significantly between pairwise comparisons of models. This observation is reasonable since passing time has business signification (e.g., waiting for approval or feedback from lawyers) that supports classification but does not add information to support early detection. We also observe that using an explicit layer to model the monthly aggregation of inputs is valuable. In other words, learning to emphasize notes for a given month is beneficial to generating the associated risk factors.

4.4.2 Comparing Our Models

We present in Table 2 the two performance metrics we used to compare models in our paper for every early detection model evaluated.

Model	F1 Macro (%)	ED (%)
Logistic	78.0 ± 0.00	69.18 ± 0.00
LSTM-M	80.26 ± 0.69	74.19 ± 0.61
SMART	79.51 ± 1.20	49.55 ± 0.62
SMARTR	**81.16** ± 0.25	**47.37** ± 0.86

Table 2: Classification Performances for Models

As we can see, Our SMARTR model outperforms the SMART model (our SAFE adaptation) and both baseline models for a posteriori and early classification. We notice that for early detection purposes, SMARTR requires, on average, 2.18 % fewer documents than SAFE to obtain correct predictions, making it roughly 4 % faster to detect expensive claims. These observations provide an obvious but essential insight that exploiting the time dimensions within a longitudinal context has significant value. We present in Figure 2 the evolution of the F1 score average and 95 % confidence interval as a function of the number of notes used for classification for each model.

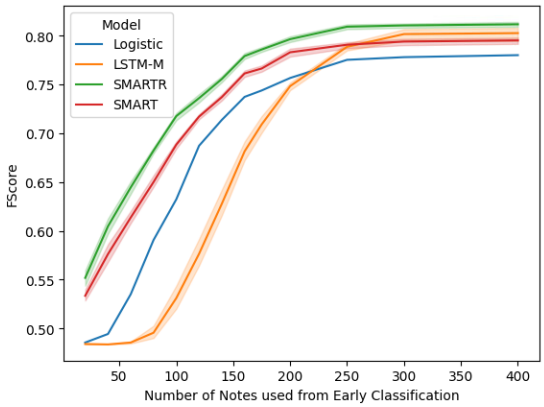

Figure 2: F1 score metric for models using a limited amount of notes

The lines on this figure are coherent with the values presented in Table 2; we can see that the green curve associated with the SMARTR model is above every other curve. Furthermore, as its confidence interval does not overlap with another curve, we can conclude that the performance of SMARTR is significantly better than SAFE and other baselines at every timestep during inference.

5 Conclusion

In this paper, we have proposed the SMARTR model and evaluated its enhancement. Our results show that our approach improves overall classification performance compared to the SAFE model and allows a 4% faster early detection. Our enhancements were tested on a longitudinal corpus comprised of claim files, where the early detection of expensive claims was the task to achieve.

Future work includes using an LLM to aggregate texts from many notes and obtain key elements of a claim or a multi-decrement approach to model the probability that the claim settles without becoming expensive. This approach would allow the model to discern the common, less costly elements of both types of claims and those associated with claims that will be closed without becoming costly.

Limitations

The main limitation of our work is that our approach could only be tested on the proprietary dataset provided for this study. This proprietary dataset contains unique characteristics but is common to datasets held by various insurance companies, so these results likely apply to them.

References

Bart Baesens, Tony Van Gestel, Maria Stepanova, Dirk Van den Poel, and Jan Vanthienen. 2005. Neural network survival analysis for personal loan data. *Journal of the Operational Research Society*, 56(9):1089–1098.

Jean-Thomas Baillargeon and Luc Lamontagne. 2024. Same file prediction: A new pretraining objective for bert-like transformers. In *Proceedings of the 37th Canadian Conference on Artificial Intelligence (in press)*.

Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Irfan Khan, Karen J Chandross, Michael J Pencina, Lawrence Carin, and Ricardo Henao. 2020. Calibration and uncertainty in neural time-to-event modeling. *IEEE transactions on neural networks and learning systems*.

David Faraggi and Richard Simon. 1995. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. 2018. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12.

Can Liu, Qiwei Zhong, Xiang Ao, Li Sun, Wangli Lin, Jinghua Feng, Qing He, and Jiayu Tang. 2020. Fraud transactions detection via behavior tree with local intention calibration. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3035–3043.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Dan Pan, An Zeng, Longfei Jia, Yin Huang, Tory Frizzell, and Xiaowei Song. 2020. Early detection of alzheimer's disease using magnetic resonance imaging: a novel approach combining convolutional neural networks and ensemble learning. *Frontiers in neuroscience*, 14:259.

Akey Sungheetha et al. 2021. Design an early detection and classification for diabetic retinopathy by deep feature extraction based convolution neural network. *Journal of Trends in Computer Science and Smart Technology*, 3(2):81–94.

Ping Wang, Yan Li, and Chandan K Reddy. 2019. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36.

Anny Xiang, Pablo Lapuerta, Alex Ryutov, Jonathan Buckley, and Stanley Azen. 2000. Comparison of the performance of neural network methods and cox regression for censored survival data. *Computational statistics & data analysis*, 34(2):243–257.

Fei Xiao, Yuncheng Wu, Meihui Zhang, Gang Chen, and Beng Chin Ooi. 2023. Mint: Detecting fraudulent behaviors from time-series relational data. *Proceedings of the VLDB Endowment*, 16(12):3610–3623.

Shuhan Yuan, Panpan Zheng, Xintao Wu, and Yang Xiang. 2017. Wikipedia vandal early detection: from user behavior to user embedding. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 17*, pages 832–846. Springer.

Panpan Zheng, Shuhan Yuan, and Xintao Wu. 2019. Safe: A neural survival analysis model for fraud early detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1278–1285.