DSO 530 Group Project 2025

1 Insurance Loss Analytics Group Project

One of the main challenges in the insurance industry is setting the right premium for policyholders. In a competitive market, insurers must base premiums on the expected loss for each policyholder. For example, if an insurance company charges older drivers too much and younger drivers too little, it could lose profitable customers as older drivers switch to other companies, leaving the insurer with underpriced policies for younger drivers. This creates an adverse selection problem, where the insurer is left with higher-risk customers, leading to financial loss. To avoid this, insurers need to accurately predict expected losses and adjust premiums accordingly.

In this project, we focus on predicting claim losses, which is key to setting the right premiums. Being able to predict these losses helps ensure fair pricing, leading to a balanced and profitable portfolio and reducing the risk of adverse selection. However, predicting these losses is challenging because claim sizes are often highly skewed, and many claims are zero (indicating no loss). This type of data is hard to transform into a normal distribution, and special treatment is needed for zero claims.

Claim loss studies involve claim data that are often assumed to follow a *Tweedie distribution* (for details see here: `https://en.wikipedia.org/wiki/Tweedie_distribution`), which can model both the frequency and severity of claims. Tweedie distributions comprise of a family of probability distributions which include the purely continuous normal, gamma and inverse Gaussian distributions, the purely discrete scaled Poisson distribution, and the class of compound Poisson–gamma distributions which have positive mass at zero, but are otherwise continuous. Data following this distribution require that popular machine learning models, such as the generalized linear model (GLM), random forests, booting, neural networks etc be adapted effectively to account for the unique characteristics of the loss function under the Tweedie distribution. In this project, we explore how to adapt these models to the *Tweedie loss* in order to predict important insurance metrics, including loss cost per exposure unit and historically adjusted loss cost. Additionally, we will conduct inference to understand the key variables that drive the successful prediction of these quantities, providing business insights into the underlying patterns and relationships within the data.

2 Available Datasets

You will find two datasets on Brightspace:

- `insurance_train.csv`: Contains data for 37451 policy transactions on variables `X.1, ..., X.28` which are described in Table 1 (on page 5).
- `insurance_test.csv`: Contains 15787 policy transactions on all variables except `X.1, X.15, X.16, X.17, X.18`.

3 Project Tasks and Goals

The core idea of the project is to use the training data to adapt the statistical/ML models we are studying in this course for the following tasks.

3.1 Task 1: Predicting Loss Cost per Exposure Unit and Historically Adjusted Loss Cost

Predict the loss cost per exposure unit (LC) and historically adjusted loss cost (HALC) using the training data for new policyholders in the test set. LC and HALC are defined as:

- **Loss Cost per Exposure Unit (LC):**

$$\text{LC} = \frac{X.15}{X.16}$$

- **Historically Adjusted Loss Cost (HALC):**

$$\text{HALC} = \frac{X.15}{X.16} \times X.18$$

3.2 Task 2: Predicting Claim Likelihood (Binary Classification)

Predict Claim Status (CS) which is 1 if the a new policyholder in the test set makes a claim, and 0 if they do not.

Ultimately you will select what you consider to be the most accurate approach and use it to make predictions for LC, HALC and CS. I will compare these predictions to their actual values on the test policyholder records (which I have), in terms of **out-of-sample MSE** for LC and HALC and **ROC-AUC** for CS, respectively.

4 Prediction Submission Guidelines

Group x must submit a CSV file `group_x_prediction.csv` containing three columns: `LC`, `HALC` and `CS`. It is required that you follow this naming convention.

5 Grading Criteria

The project will be graded out of 13 points (see Table 2 on page 6). In addition, there is an opportunity for getting **2 bonus points for being more creative**. 7 points will be allocated to the project report and 6 points will be allocated to the presentation stage. **Every member of the group should speak in the presentation. And the video of the speakers should appear in the recording.**

5.1 Instructions for write-up

You will submit a report documenting your analysis. A typical format for the report would be:

- One cover page and one Executive Summary page.
- Review of the approaches that you tried or thought about trying.
- Summary of the final approaches that you used to predict LC, HALC and CS, and why you chose those approaches, detailing the model selection steps.
- Discuss model interpretation; you might consider using SHAP values for this.
- Highlight clearly the part that is focused on your innovations beyond the questions discussed here. This will be considered for awarding the bonus points.
- Conclusions.

Among other things, points will be allocated for clear articulation of the approaches you considered, and the reason you chose the final approaches. **The main body of the report must be no longer than six pages** (including the the cover page and the Executive Summary page). However, you may also include a (up to) thirty page technical appendix with various computer outputs to justify the conclusions in your report. I will mainly look at the six-page main body, but might refer to the appendix if I see something unreasonable. The report should be named as **group_x_report.pdf**, for group x. All group members' names and student id numbers should be clearly indicated on the cover page. Also, please include the contact person's email (**and that person's email only**).

5.2 Business understandings

You should consider thinking about the business problem at hand. Here are some key aspects to consider, though this list is only a starting point:

1. in the above prediction problems, would you argue if prediction accuracy or interpretation is more important? Why?
2. what are the key challenges faced by machine learning models in this context?
3. can you argue from a business perspective which predictor variables should be included in your prediction (i.e., is variable selection necessary)?
4. Are you comfortable about directly using your trained model to predict loss cost for life insurance policies? Why?

5.3 Presentation and student voting scheme

Only the top ranked videos will be played in class. I will randomly divide all groups in five cohorts (I, II, III, IV and V). Each group will watch all group presentation videos of another cohort. For example, each group in cohort I will watch all videos in cohort II. The contact person of each group should email to the TA their choices of top 3 (we do not distinguish among the top three) by **5 pm on April 28. The email should be titled: "group x round 1 vote" (x indicates your group number).** In this round, the voting criteria include clarity of the presentation, rigor of the approaches, and creativity. The TA will aggregate the votes and announce the top three in each cohort. So there will be fifteen groups selected in the first round. **The class will be informed the 15 finalists on April 29.**

On **Wednesday, April 30**, in each class, we will play videos of four selected groups (one of the sections will have three videos), followed by my comments and Q&A. As we have four sessions, in total, I will play 15 group presentation videos on that day. **Attendance on that day is mandatory.** Round 2 will consist of selection of the top 5 teams of the 15 finalists by the instructor Prof. Paromita Dubey to determine which teams receive the full 6 / 6 presentation points. The remaining teams will get a maximum of 5 out of 6 presentation points. In case there are ties, the TA votes will be the tie breakers. **The results of round 2 votes will be announced on May 5.**

If your group does not participate in the voting, your group will be removed from other groups' votes (if there are any) and up to **1 point** will be taken away from your presentation score. **No group should lobby for votes.**

6 Deliverables (no late submission accepted)

1. Upload your **(up to) 10 min presentation video in mp4 format (group_x_presentation.mp4)**. Deadline: **April 26 at 11:59 pm**. The **contact person** should submit your group video to: `https://www.dropbox.com/request/APueboB4jYwv1G85gtjb` . Please double-check your video before uploading it. **A video longer than 10 min faces penalty.**
2. Submit two files, i.e., your report (**group_x_report.pdf**) and prediction (**group_x_prediction.csv**). Deadline: **May 8 at 5 pm**. The **contact person** should submit to: `https://www.dropbox.com/request/U4qE7KuOpMPJdjydGn3a`

Additional instruction for Dropbox submission: If you are not signed in with an email account on the browser, you will be asked to give your first name, last name and email address. If you already signed in with some email account, its information will be used automatically. **After the deadline, the link will be deactivated.** If you submit multiple times before the deadline, the last submission (with the same file name) will overwrite previous submissions.

7 Timeline

1. April 26 at 11:59 pm, video upload
2. April 28 at 5 pm, round 1 vote

3. April 29, round 1 finalists will be announced
4. April 30 in class, round 1 finalists' video play, Q&A and comments. Attendence is required on April 30.
5. May 5, round 2 results
6. May 8 at 5 pm, report and prediction

8 Group meeting with the instructor

After May 4, the other groups (minus the 15 groups whose videos are commented in class) has an option to meet with the instructor for short consultations on specific aspects of their project. A schedule will be given later.

9 Optional

If you prefer a different topic than the one proposed, you are welcome to select your own. Please prepare a project proposal outlining the main tasks you plan to address and submit it to your instructor, Prof. Paromita Dubey, by **Thursday, April 3**. The proposed topic should be of similar difficulty to the one outlined above and must be approved by your instructor. Once your proposal is received, your instructor will schedule an appointment with you to let you know if your project proposal was accepted and finalize the next steps.

If you are interested in a more challenging project, you may consider participating in active Kaggle competitions. The intent to pursue this as a group project must also be confirmed by **Thursday, April 3**.

Variable	Description
X.1	Internal identification number assigned to each annual contract formalized by an insured.
X.2	Start date of the policyholder's contract (DD/MM/YYYY).
X.3	Date of last contract renewal (DD/MM/YYYY).
X.4	Date of the next contract renewal (DD/MM/YYYY).
X.5	Date of birth of the insured declared in the policy (DD/MM/YYYY).
X.6	Date of issuance of the insured person's driver's license (DD/MM/YYYY).
X.7	Channel through which the policy was contracted (0: Agent, 1: Insurance brokers).
X.8	Total number of years that the insured has been associated with the insurance entity.
X.9	Total number of policies held by the insured in the insurance entity.
X.10	Maximum number of policies that the insured has ever had in force.
X.11	Maximum number of products that the insured has simultaneously held at any given point.
X.12	Number of policies canceled or terminated for nonpayment in the current year.
X.13	Last payment method of the reference policy (1: half-yearly, 0: annual).
X.14	Net premium amount associated with the policy during the current year.
X.15	Total cost of claims for the insurance policy during the current year.
X.16	Total number of claims incurred for the insurance policy during the current year.
X.17	Total number of claims filed throughout the entire duration of the policy.
X.18	Ratio of the number of claims filed to the total duration (years) of the policy in force.
X.19	Type of risk (1: motorbikes, 2: vans, 3: passenger cars, 4: agricultural vehicles).
X.20	0 for rural, 1 for urban (more than 30,000 inhabitants).
X.21	1 if multiple regular drivers are declared, 0 if only one driver is declared.
X.22	Year of vehicle registration (YYYY).
X.23	Vehicle power measured in horsepower.
X.24	Cylinder capacity of the vehicle.
X.25	Market value of the vehicle as of 31/12/2019.
X.26	Number of vehicle doors.
X.27	Energy source used to power the vehicle (P: Petrol, D: Diesel).
X.28	Vehicle weight in kilograms.

Table 1: Description of Selected Variables

Section	**Points**	**Evaluation Criteria**
Project Report Write-up	3	See above for further instructions.
Task 1 predictive performance	2	0 points will be awarded to groups whose efforts do not meet expectations. 1 point will be awarded to groups that show significant effort, but whose model performance is considerably below that of other groups. 2 points will be awarded to groups whose model performance is comparable to that of their peers.
Task 2 predictive performance	2	0 points will be awarded to groups whose efforts do not meet expectations. 1 point will be awarded to groups that show significant effort, but whose model performance is considerably below that of other groups. 2 points will be awarded to groups whose model performance is comparable to that of their peers.
Presentation	6	6 is reserved for the top five groups that survive two rounds of popular votes (see the next page for details); 5 excellent; 4 very good; 3 good; 1-2 below the bar. Note: Voting is meant to facilitate team building (and fun). Teams are encouraged to inspire each other in creating better strategies. Points will be subtracted for teams not participating in the voting.
Bonus	2	The instructor might award bonus points based on the overall creativity demonstrated by the group, including formulating innovative questions beyond what is described and designing approaches to solve them.
Total	13 + (2 bonus points)	

Table 2: Project Grading Rubric Breakdown