ABDA PROJECT PROPOSAL

Roshni Kukreja

Santhosh Sakthyavinayagam

December 2024

1 Dataset Information

1.1 Source

Link : https://data.mendeley.com/datasets/992mh7dk9y/2

Publisher: Mendley Datasets Dimension : 1000 x 40

1.2 Data Description

Our data describes 1000 records of insurance claims made in the US in 2015. We summarize the relevant feature of our dataset in Table 1.

Category **Description**	**Feature**
Demographics Age of the policyholder	`age`
Gender of the insured individual	`insured_sex`
Education level of the insured individual	`insured_education_level`
Occupation of the policyholder	`insured_occupation`
Relationship status	`insured_relationship`
Policy Details Unique policy identifier	`policy_number`
State where the policy was issued	`policy_state`
Coverage limits (e.g., 250/500)	`policy_csl`

Category **Description**	**Feature**
Deductible amount	`policy_deductable`
Annual premium amount	`policy_annual_premium`
Claims Information Total amount claimed	`total_claim_amount`
Amount claimed for injuries	`injury_claim`
Amount claimed for property damage	`property_claim`
Amount claimed for vehicle damage	`vehicle_claim`
Fraud indicator (Y/N)	`fraud_reported`
Incident Details Type of incident (e.g., collision, theft)	`incident_type`
Severity of the incident	`incident_severity`
Authorities contacted	`authorities_contacted`
Time of the incident	`incident_hour_of_the_day`
Other Variables Make of the vehicle	`auto_make`
Model of the vehicle	`auto_model`
Manufacturing year of the vehicle	`auto_year`

Table 1: Dataset Feature Description

2 Project Objective and Respective Plans

2.1 Objective Idea 1

Objective 1 : Predict the probability of insurance fraud (`fraud_reported` as in the dataset) and identify high-risk claims for auto insurance.

Our starting point is to apply the Bayesian Logistic Regression Model because we have a binary target and logistic regression likelihood is appropriate for binary classification (Equation 1).

$$P(y_i = 1 \mid \mathbf{x}_i, \beta) = \frac{1}{1 + \exp\left(-(\beta_0 + \mathbf{x}_i\beta)\right)} \tag{1}$$

From our initial research, we did not come across many papers that had this approach. Except for research carried out on accidents that occurred in Massachusetts, USA during 1993 by using Bayesian Neural networks [2].

2.2 Objective Idea 2

Objective 2 : Our second potential project objective is to estimate the claim risk and determine optimal insurance premiums for motor vehicle insurance claims using Bayesian methods.

By leveraging the provided dataset, the project will aim to model the frequency and severity of insurance claims using appropriate Bayesian probabilistic models, such as the Poisson distribution for claim frequencies and the Gamma distribution for claim amounts, inspired by the methodology used in the reference paper by Sukono et al. (2018) [1].

References

[1] Sukono et al. “Model estimation of claim risk and premium for motor vehicle insurance by using Bayesian method”. In: *IOP Conference Series: Materials Science and Engineering* 300.1 (2018), p. 012027. DOI: 10.1088/1757-899X/300/1/012027.

[2] G. I. Webb. “Opus: An efficient admissible algorithm for unordered search”. In: *Expert Systems with Applications* 28.2 (2005), pp. 365–377. DOI: 10.1016/j.eswa.2005.04.030.