Name: Chaitanya Darade

Homework-3

Individual Assignment on Project Understanding

Business and Data Understanding

All insurance companies will have a loss ratio, as it is unlikely that all policyholders will fail to make a claim across the period of their insurance, otherwise insurance would be unnecessary and there would be no insurance companies. A high loss ratio—meaning one where the losses approach, equal, or exceed the premium—is considered bad. To minimize this loss ratio, insurers develop estimates as to how many claims they may expect to see and how expensive the claims will be based on the types of policies they provide to policyholders.

So, this Loss ratio is calculated by below formulas:

Loss_ratio = Loss_Amount/Annual_Premium

Where,

Loss_amount = Severity * Claim_Count

With a supervised machine learning strategy, loss ratio would be considered as a target variable and will take continuous values. By finding and analyzing patterns for high loss ratio, the policymaker can take several steps to reduce loss ratio e.g. by recommending a higher deductible, i.e. you lower the amount of premium you receive. But that doesn't matter because the potential payouts keep your loss ratio a lot lower or Secure referrals from clients you know have excellent records. Most people enjoy the company of people similar to them. The maniac driver is hardly included in the carpool-driving list. Asking your clients for names of people they consider good drivers is safe and effective approach.

Data Preparation

Two approaches are available to prepare the training dataset:

1. **Naïve Approach** :

    - Use the training dataset to build a model that given policy attributes predicts the loss of that individual policy
    - Use the above model to predict the loss of all policies in a testing portfolio
    - Once the losses are available, now you can compute the natural logarithm of the loss ratio of the entire test portfolio

2. **Better Approach** :

    - Create training portfolios similarly as testing portfolios.
    - For each training portfolio engineer a set of features that summarizes the data in that portfolio i.e. mean driver age etc.
    - Create a new training dataset using these new features and ln_LR and train a model
    - Use these features and the model to predict the testing portfolios

To form the training dataset, the second approach is better to avoid high computation cost in calculating ln_LR for each policy as we have a huge amount of input data. Also, it synchronizes with the testing dataset format, which would be more feasible for the test process.
All missing values would be replaced or imputed using different strategies depending on the business needs.

Modelling

The regression model would be used to calculate weights on training dataset and predict the loss ratio on the testing set. Each portfolio will have more than 1000 policies, whose relevant features are summarized and 1 record per portfolio would be created. In a way, the final dataset will contain a number of rows equal to number of portfolios we have created, and we train the regression model on it. This trained model will predict ln_LR on the testing dataset.

To form the training dataset, this approach is taken to avoid high computation costs in calculating ln_LR for each policy as we have a huge amount of input data. Also, it synchronizes with the testing dataset format, which would give more accurate results.

The choice of linear model seems appropriate for this model, as the target variable loss ratio is calculated using loss divided by annual premium amount, and loss is calculated, severity*claim count. These values seem linearly dependent on some other attributes such as vehicle manufacturing year, vehicle performance, vehicle usage, anti-theft service, driver's age, vehicle driver points**,** etc., however, in our testing dataset don't have few decisive columns(claim_count, loss-amount, Frequency, Severity), first model need to be trained to calculate these values first also. To do that finding features on which these values are dependent is necessary and use these trained models to calculate these column values for testing dataset as well. So, basically three

models are required to calculate claim_count, severity and finally the loss_ratio to generalize on the test data. All are required to find appropriate features, which gives hindsight of the pattern for loss amount and number of claims, will automatically calculate correct weights and predict loss ratio accurately on testing dataset.

Evaluation and Deployment

After the model training, its vital to evaluate the model's performance using different measures such as cross-validation, AUC_RUC curve, etc. to evaluate how well this model generalizes on real-world data. To examine this scenario, an expert's opinion or testing is mandatory as he holds through business knowledge. The Business team will understand the results, i.e. loss ratio per portfolio, by considering factors such as cost, benefit, the threshold was chosen etc. Once they find some insights for example. about the vehicle/driver related patterns who are making a large number of claims, etc. and taking corrective action in policy building, the model would be deployed on real-world data. So huge real-time datasets would then be evaluated against this model and steps to reduce loss ratio as much as possible can be taken.

Following evaluations/actions can be taken by policymakers:

1. A company that had a few major claims may have an inflated loss ratio but isn't truly a risky client to retain. For those companies, insurers can assist in implementing strategies that prevent severe losses in the future.

2. Another significant contributor to high loss ratios is fraud. Fraud is difficult to investigate effectively, but paying for fake or inflated claims eats into profits. Automating fraud detection can help expedite the process since all relevant data and cases are centralized. Automation can set up rules and alerts to identify cases with a high chance of being fraudulent. Automation can also help insurance providers understand a claim's path with predictive modeling. Machines can process data from prior claims to provide insight into future ones

3. By recommending a higher deductible, you lower the amount of premium you receive. But that doesn't matter because the potential payouts keep your loss ratio a lot lower.

4. Secure referrals from clients you know have excellent records. Most people enjoy the company of people similar to them. The maniac driver is hardly included in the carpool-driving list. Ask your clients for names of people they consider good drivers. Explain that the fewer accidents paid by the company, the lower the premium becomes in the future.