Predicting the Car Insurance Claim Amount: Using Subset Selection and Cross-Validation

Yihan Zhang, Yifei Chen

May 10, 2025

1 Introduction

As of 2022—the most recent year for which the Insurance Research Council collected national data—approximately 86% of drivers in the U.S. had car insurance (Autoinsurance, 2024). Drivers rely on it for a sense of security, helping to safeguard against financial risk and unexpected losses.

One of the key concerns for policyholders after an accident is the amount of compensation they may receive. This payout often depends on multiple factors, including the driver's education level, the type of incident, and whether property damage was involved. Accurately predicting the total claim amount based on these features can help insurance companies improve risk assessment and develop more personalized premium structures. From a consumer perspective, such insights also promote transparency and fairness in the claims process.

In this paper, we aim to predict the total claim amount based on such relevant variables. We will begin by using backward subset selection to identify a predictive model. Then, we will randomly select two models and evaluate their performance through cross-validation (CV).

2 Data Preprocessing

We accessed the dataset "Insurance Fraud Claims Detection" from GitHub (2018). The GitHub repository lacks detailed documentation regarding the data's provenance and year of collection, which raises some concerns about the reliability and authenticity of the dataset.

This dataset contains 1000 rows, where each observation represents an insurance claim. It includes basic information about the policyholder, details of the incident, and the claim amount, comprising a total of 35 variables. We manually selected six quantitative variables and six categorical variables as predictors for our models, with `total_claim_amount` serving as the outcome variable.

2.1 Selected Variables

Table 1: Selected variables and their descriptions

Variable	Type	Description
Months as customer	Quantitative	Number of months that the person has been a customer (month)
Age	Quantitative	The age of the customer (year)
Number of vehicles involved	Quantitative	How many vehicles were involved in this accident (number of cars)
Bodily injuries	Quantitative	Number of bodily injuries in the accident (number of people)
Injury claim	Quantitative	Amount claimed for injuries ($)
Property claim	Quantitative	Amount claimed for property damage ($)
Insured sex	Categorical	Gender (Male/Female)
Property damage	Categorical	Whether property damage occurred (Yes/No)
Edu level	Categorical	Education level (High School, Associates, College, Advanced Degree)
Incident type	Categorical	Type of incident (Multi-Vehicle Collision, Parked car, etc.)
Incident severity	Categorical	Severity (major damage, minor damage, etc.)
Total claim amount	Quantitative	Response variable ($)

3 Data Cleaning

We began by renaming two variables to make them more convenient for use in code. In the `edu_level` category, we consolidated Masters, JD, MD, and PhD into a single category labeled "advanced degree." We also ensured that both quantitative and categorical variables were correctly converted to appropriate data types for modeling purposes.

We identified 360 missing values in the `property_damage` variable. After examining the distribution of missing values across different categories, we did not observe any clear pattern. Furthermore, the data source did not provide any explanation for the missing entries. As a result, we chose to remove these observations from the dataset. After this cleaning step, we retained 640 complete observation units for analysis.

4 Exploratory Data Analysis

We conducted an exploratory data analysis, during which several noteworthy observations emerged.

4.1 Distribution of Claims

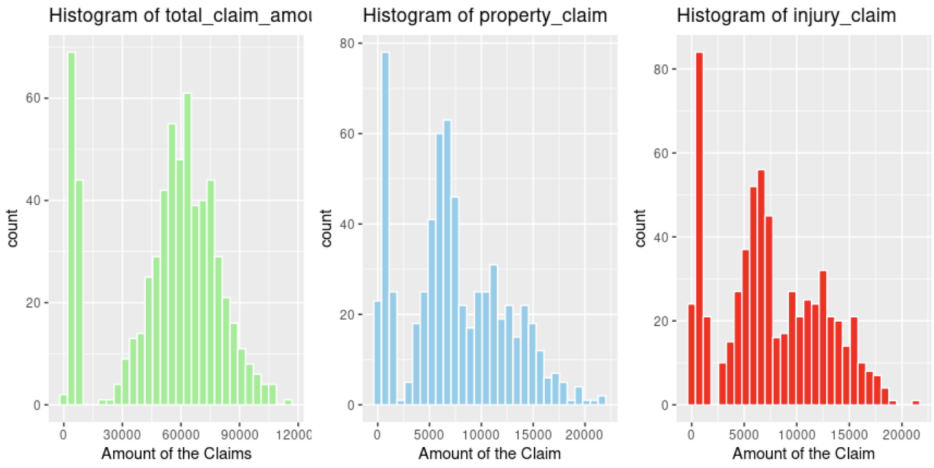

Figure 1: Histogram of claim amounts

A histogram of the outcome variable `total_claim_amount` revealed that approximately 18.75% of the claims are relatively low, concentrated in the \$0–10,000 range. The remaining claims appear to follow a roughly normal distribution, centered around \$60,000, with a range spanning from \$30,000 to \$100,000. The distributions of `property_claim` and `injury_claim` are quite similar: both also exhibit a spike near zero representing 18.75% of the data, followed by a right-skewed distribution for the remaining values.

4.2 Education Level Analysis

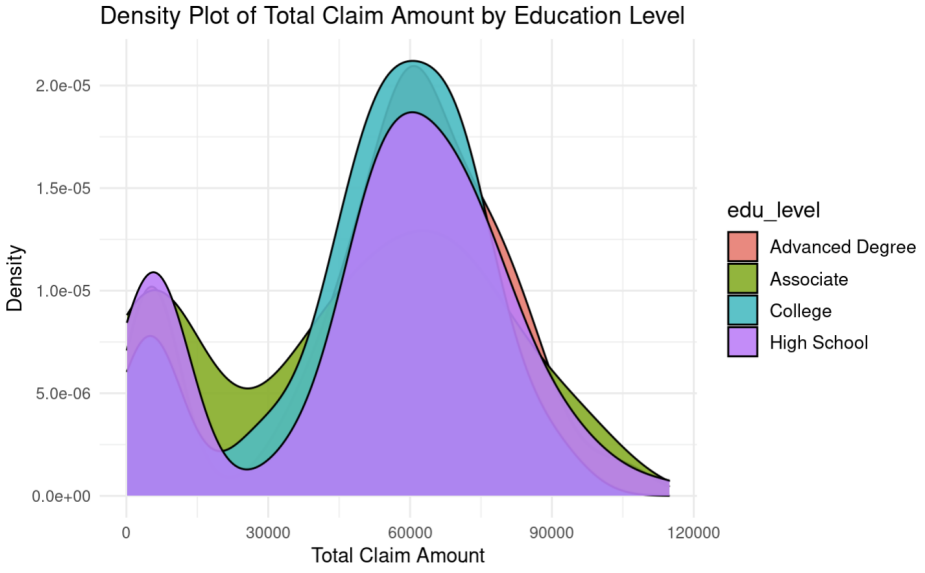

Figure 2: Density plot by education level showing differential peaks

A density plot was constructed with `total_claim_amount` on the x-axis, color-coded by different education levels. The plot revealed that all education levels exhibit a small peak near zero and a prominent peak around $60,000. Notably, individuals with college or advanced degrees showed a slightly more pronounced peak at $60,000, which might suggest they possess better negotiation skills that help them secure higher claim amounts.

4.3 Fraud Detection

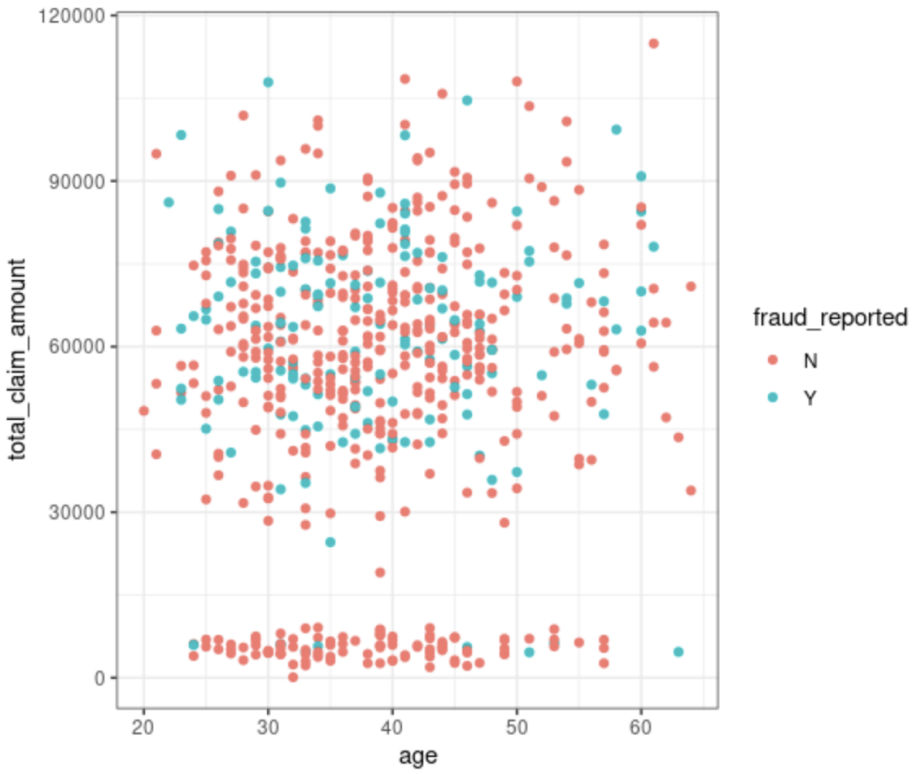

Figure 3: Fraud frequency by claim amount range

Consistent with our expectations, we observed that nearly all low-amount claims were not fraudulent, while more than 50% of high-amount claims were flagged as fraud, indicating a potential link between claim size and fraudulent behavior.

4.4 Added Variable Plots

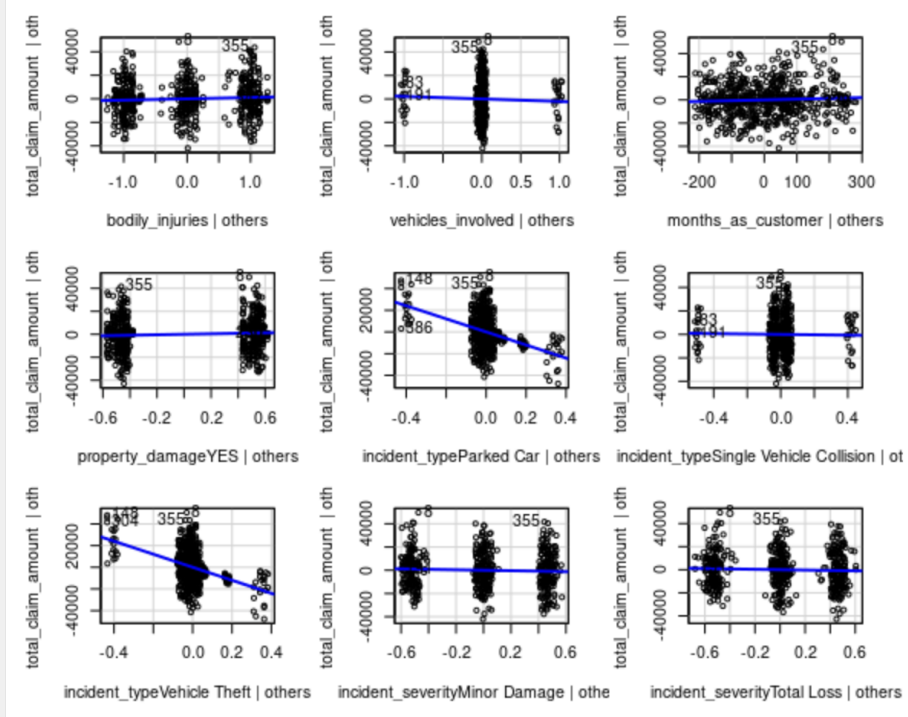

Figure 4: Added-variable plots for key predictors

Last but not least, we used added variable plots to examine whether individual predictors exert a significant influence on `total_claim_amount` after controlling for other variables. From these plots, we found that incidents categorized as vehicle theft or parked car tend to have a negative effect on the total claim amount. Additionally, number of years a customer has been with the company shows a positive association with `total_claim_amount`, and a weak positive relationship between the number of bodily injuries reported in the claim and the `total_claim_amount`. Last but not least...

5 Model Prediction & Evaluation

5.1 Backward Subset Selection

Backward subset selection is a variable selection procedure that identifies the best subset of predictors for each model size. Specifically, for each number of predictors k, it evaluates all possible models that contain k predictors and selects the one with the highest R^2 on the training data, denoted as M_k. This results in a sequence of models $M_1, M_2, \ldots, M_n$, where n is the total number of available predictors. To evaluate the out-of-sample performance of these models, we then apply 10-fold cross-validation to estimate the test Mean Squared Error (MSE) for each M_k. The model with the lowest cross-validated test MSE is selected as the best-performing model. According to

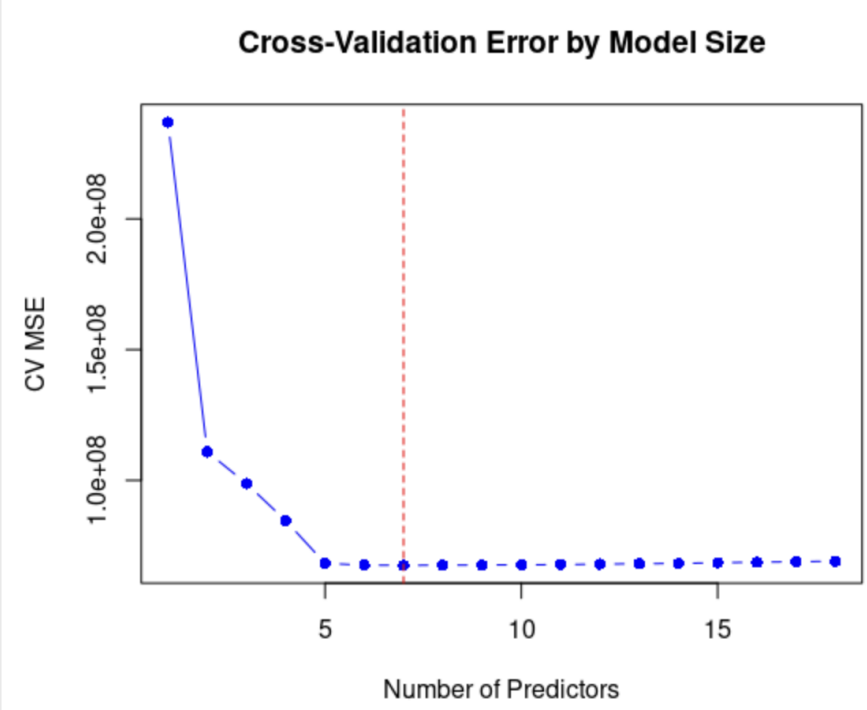

Figure 5: Number of Predictors v. MSE

the results, the model with $k = 7$ predictors achieved the lowest average test MSE of 67,466,469, and is therefore selected as the best-performing model. The predictors included in this model are:

- `injury_claim`
- `property_claim`
- `edu_levelCollege`
- `incident_typeMulti-vehicle Collision`
- `incident_typeParkedCar`
- `incident_typeSingle Vehicle Collision`
- `fraud_reportedN`

Note that backward selection treats each level of a categorical variable as a separate dummy variable. This is why multiple levels of `incident_type` appear as separate predictors in the final model.

5.2 Cross-Validation Comparison

Cross-validation (CV) serves three primary purposes: (1) tuning parameter selection, (2) variable selection, and (3) evaluating model performance to avoid overfitting.

In k-fold cross-validation, the dataset is randomly partitioned into k disjoint subsets (folds) of approximately equal size. For each fold $j = 1, 2, \ldots, k$, the model is trained on the data excluding the j-th fold, and its prediction error is evaluated on the held-out fold. The estimated mean squared error (MSE) is calculated as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i) \right)^2 \tag{1}$$

The cross-validation error is then:

$$\text{CV} = \frac{1}{k} \sum_{i=1}^{k} \text{MSE}_i \tag{2}$$

We randomly selected two models for comparison:

- **Model A**: A simple linear model: `total_claim_amount` $\sim$ `age + vehicles_involved + injury_claim`
- **Model B**: A full linear model with all predictors: `total_claim_amount` $\sim$ `months_as_customer + age + vehicles_involved + bodily_injuries + injury_claim + property_claim + insured_sex + property_damage + edu_level + incident_type + incident_severity + fraud_reported`

We applied 10-fold cross-validation ($k = 10$), where the dataset is divided into 10 folds. Each fold takes turns serving as the test set, while the remaining nine folds form the training set. We chose MSE as the error metric because it heavily penalizes large prediction errors, making it sensitive to outliers.

The output metrics compared for both models are:

- `CV_MSE_Mean`: Average test MSE across folds
- `CV_MSE_SD`: Standard deviation of test MSE across folds
- `InSample_MSE`: Training MSE from the full dataset

Model <chr>	CV_MSE_Mean <dbl>	CV_MSE_SD <dbl>	InSample_MSE <dbl>	InSample_SD <dbl>
Model A	233767567	40917704	229954703	322843034
Model B	69081551	11571068	65065375	107280574

Figure 6: Comparison of Model A and Model B performance when k = 10

The results showed that Model B consistently had lower values across all three metrics compared to Model A under 10-fold CV, indicating that Model B provides a better fit to the data.

Additionally, we explored different values of k, including leave-one-out cross-validation (LOOCV, where $k = n - 1$), and $k = 5$. We observed that the in-sample MSE for both Model A and Model B remained the same across all values of k, which makes sense because in-sample MSE is calculated

using the entire training data and is unaffected by how we split the data for cross-validation. Interestingly, for Model A, the CV_MSE_Mean was slightly lower under LOOCV compared to $k = 10$, while for Model B, it was slightly higher. Additionally, we found that CV_MSE_SD was lowest for both models when $k = 5$, likely because fewer folds reduce the risk of overfitting to small validation sets, thus lowering the variance in prediction error.

Model <chr>	CV_MSE_Mean <dbl>	CV_MSE_SD <dbl>	InSample_MSE <dbl>	InSample_SD <dbl>
Model A	232967258	327286923	229954703	322843034
Model B	68867218	113618315	65065375	107280574

Figure 7: k = n - 1 (639)

Model <chr>	CV_MSE_Mean <dbl>	CV_MSE_SD <dbl>	InSample_MSE <dbl>	InSample_SD <dbl>
Model A	233223160	19807450	229954703	322843034
Model B	68274384	8752033	65065375	107280574

Figure 8: k = 5

Taking all this into account, we would choose Model B with $k = 5$, because even though its CV_MSE_Mean is slightly higher than in LOOCV and $k = 10$, it offers the lowest variance (CV_MSE_SD), indicating a more stable and generalizable model.

6 Conclusion

In conclusion, we ran the backward selection model and showed that the model with 7 predictors works best. This paper then compared the performance of two models with different numbers of folds using cross-validation, with Model B consistently outperforming Model A across all folds.

To achieve more accurate predictions (in terms of MSE and variance), it is necessary to incorporate more predictors that are related to `total_claim_amount` and utilize a larger dataset. So, the next step will be to explore additional variables, refine feature selection, and potentially gather more data to improve model performance.

7 Citations

References

[1] AutoInsurance.com (2024) *Uninsured motorists statistics: 2024 U.S. report.* Available at: `https://www.autoinsurance.com/research/uninsured-motorists/` (Accessed: 3 May 2025).

[2] Cahalan, C. (2024) *How many car insurance claims are filed each year? 2025*, ConsumerAffairs. Available at: `https://www.consumeraffairs.com/insurance/car-insurance-claims-statistics.html` (Accessed: 3 May 2025).

[3] Mwitiderrick (n.d.) *Insurance_claims data*, GitHub. Available at: `https://github.com/mwitiderrick/insurancedata` (Accessed: 3 May 2025).