Data Analysis Report: Predictive Modelling of Insurance Claim Amounts

Abstract

In this report, we address the problem of predictive modelling of auto insurance claim amounts through structured claim data using supervised machine learning. Following extensive data pre-processing that consisted of cleaning, processing of missing values, and encoding, five regression models were evaluated: Linear Regression, Ridge, Lasso, Elastic Net, and Random Forest. We measured the performance of the models with R^2, adjusted R^2 and AIC. The linear and regularized models had limited predictive performance (test Adjusted R^2 ~3–4%), but the Random Forest could achieve slightly higher performance (test Adjusted R^2 ~8.9%) being able to capture non-linear relations. Important features and SHAP analysis identified the important cost factors as policy length, time to report, and severity. The report also identifies modelling limitations and suggests future development through enhanced algorithms and better data sets.

Table of Contents

1. Introduction

Precise estimation of the amount of an insurance claim is important for insurers to price policies, to control risk and to set aside reserves properly. Claims data in auto insurance usually contain structured information describing details of the policyholder, vehicle, the circumstances of an

incident and other involved third-party entity. Previously claim severity had been modelled using methods such as Generalized Linear Models (GLMs), with distributional assumption made a priori, such as Gamma or Log-normal distributions (Nelder & Wedderburn 1972). Sense and simplicity edition Consequently, GLMs provide quantitative interpretability, but they are not capable of capturing complex, non-linear interactions typically found in real-life insurance data.

We present novel approaches relying on state-of-the-art machine learning methods, which are more flexible and resilient. Ensemble methods such as Random Forests (Breiman, 2001) and boosting algorithms (Friedman, 2001) have been shown to perform better on unbalanced distributions as well as on extreme claims among others by better handling outliers and feature interactions over traditional approaches in (Henckaerts et al., 2018; Staudt & Wagner, 2021).

Our goal is to predict the cost of claims as represented in a structured dataset of auto-insurance claims. The primary goals are: 1. to compare the predictive power of several regression models; 2. to identify important predictors of claim cost levels; and 3. to investigate challenges regarding limited data and future direction.

Five types of regressions are tested: Multiple Linear, Ridge (Hoerl & Kennard, 1970), Lasso (Tibshirani, 1996) and Elastic Net and Random Forest. For linear models, we evaluate on R^2, and AIC. To improve interpretability, SHAP (SHapley Additive exPlanation) Values (Lundberg & Lee, 2017) are applied to measure feature contributions.

2. Data Preparation and Preprocessing

Prior to the data are ready for predictive modelling a lot of work in terms of data cleaning and preprocessing have been done. The raw insurance claim dataset includes 7,691 records associated to 43 features that cover policy, incident and claim details. These are across three feature types (numeric, categorical, and binary indicators).

Among numerical variables, major temporal features were covered, i.e. Inception to loss (days from policy start to incident), Notification period (delay of claim report). These time factors may affect claim severity; a longer time to report may be related to more complicated or serious cases.

Qualitative type features Notifier (a person who notifie the complain), Loss_code, Location_of_incident, and Weather_conditions describe nature of incident. These factors are categorical with many levels, which are processed using one-hot encoding (to cast them to binary features for modeling).

Binary indicators were also made available for specific conditions, e.g. TP_injury_fatality, TP_type_driver and TP_region_A to TP_region_M already in a 0/1 format for direct input in the model.

The dependent variable, Incurred, is the total claim cost (£). At first it was in a mixture of formats: symbols, such as "£", and commas, hyphens and blank entries. For simplicity, we eliminated any non-numeric information and used blanks or (probe to runway end) to denote zero. Following sanitisation, the entries were converted to floating-point format and any invalid or negative ones were removed. This converted 1754 undefined submissions to zero payouts (i.e., no payout).

There was little overall missing data, which was mainly constrained to the Weather_conditions variable (missing in ~4.5% of the records). These were imputed with a new level "Unknown" to not loose information about subjects but avoid implacable bias.

To maintain uniformity, minor modifications on column names were performed which includes the renaming of Vechile_registration_present to Vehicle_registration_present. After pre-processing, the data set had 56 features through one-hot encoding to multi-level categorical features. We dropped one category for each encoded feature in order to avoid dummy variable traps.

There was no need for feature scaling: it was not used for the models that have been chosen. Tree-based models such as Random Forest are scale-insensitive, you can use them without scaling, and with linear models such as Ridge and Lasso you have the intercept.

Lastly, the information was randomly divided into training (80%) and test (20%) using a set seed for replication. The target variable was highly skewed with a heavy tail: the median claim was ~£856, but the maximum claim was over £1.34 million. This imbalance introduces modelling difficulties, since usual regressors may underpredict not frequent, high-cost claims. These properties were taken into account during the selection and validation of models.

3. Modelling Methodology

After pre-processing, several supervised regression model-based techniques were developed to predict the amount of insurance claim. The modelling approach took a stepwise form; starting with a baseline linear regression, and sequentially progressing towards advanced techniques to potentially alleviate non-linearity, multicollinearity, and overfitting. All were trained and validated in a consistent manner and evaluated with the same set of metrics.

3.1 Modelling Techniques

Multiple Linear Regression: We first fit the benchmark model that is standard multiple linear regression based on ordinary least squares estimation. This model assumes that the relationship between the predictors and the dependent variable is linear, making it easy to interpret coefficients as marginal effects. Although easy to understand, MLR is incapable of taking into account interaction effects among features or nonlinearity. Furthermore, when many predictors are added, particularly if they are correlated with one another, the model may suffer from overfitting or instability.

Ridge Regression: Ridge Regression is introduced to solve multicollinearity and overfitting by imposing L2 penalty on the square of the coefficients (Hoerl & Kennard 1970). This approach drives the ineffective features towards zero obtaining better generalization, with the inclusion of all features. The value of regularization parameter (alpha) was initially experimented with slack regularization on relatively few variations and was chosen such that there was a balance between bias and variance without biasing too much to the latter.

Lasso Regression: Lasso Regression (Tibshirani, 1996) includes an L1 penalty that promotes sparsity by setting some coefficients to zero. This is akin to achieving embedded feature selection, leading to a simpler and possibly more understandable model. Lasso can have an edge over Ridge in the presence of irrelevant or weak predictors and particularly when the number of predictors are high. An appropriate value of alpha was selected to balance between underfitting and model sparsity.

Elastic Net Regression: Elastic Net solves the problem that arises with Ridge and Lasso by penalizing both Ridge and Lasso by the sum-of-squares to achieves shrinkage. It is especially useful when features are correlated as it can cluster them instead of picking one randomly. The model depends on 2 hyperparameters: the overall regularization strength and the L1/L2 mixing rate. Elastic Net was evaluated with a well-balanced penalty mix in this work, yet the small number of hyperparameters tuned may have led to its devitalized performance as further addressed.

Random Forest Regression: A Random Forest Regressor (Breiman, 2001) was used to accommodate the non-linearity and the complex interactions in the data. This ensemble approach combines the predictions of a number of decision trees, each one grown on a bootstrap sample and a random subset of features. Random Forests are insensitive to outliers, can handle arbitrary heterogeneous effect of features, and works pretty well on tabular data. The model was tuned with 100 estimators and the default hyperparameters, so no maximum depth constraint was set to promote full tree growth and feature interactions.

3.2 Learning and validation of the model

All models were trained on 80% of the data with the other 20% held out for out-of-sample validation. Tuning of hyperparameters (e.g., alpha for Ridge and Lasso) was performed at a high-level, exhaustive cross-validation was not. For Random Forests, hyperparameters including the depth of tree and the minimum number of samples per split were kept as default to prevent overfitting via manual tuning. Especially, we placed the test set that we step was used only for final evaluation, with all performance metrics.

3.3 Evaluation Metrics

We evaluated model performance using three primary metrics:

R^2 (Coefficient of Determination): It represents the amount of variance in the target variable explained by the model. The closer to 1 this value is the better the fit, but if you've got a high training R^2 and a low-test R^2, the model was probably over-fit.

Adjusted R^2: This takes the R^2 value and adjusts it for the number of predictors in the model: fairer comparison between models that have different numbers of predictors. It is particularly powerful for assessing the generalization to new data.

Akaike Information Criterion (AIC): Used for linear models only, it is a measure of how well the model fits the data after controlling for the number of predictors included. A lower AIC value represents more parsimonious and efficient models. Since Random Forests do not depend on likelihood-based estimation, AIC is not relevant.

Collectively, these measures offered a sound basis on which to compare model performance and inform the choice for the optimal predictor.

4. Result

Using these metrics, we compare the models' performance on the test set. Table 1 below summarizes the results, including both training and test metrics for completeness.

Table 1. Model Performance Comparison on Insurance Claims Dataset. (Train and Test results, with R^2 and Adjusted R^2 in percentage points)

Model	Train R^2 (%)	Train Adjusted R^2 (%)	Test R^2 (%)	Test Adjusted R^2 (%)	AIC (Train)

Linear Regression	20.3	19.5	6.7	3.1	142386
Ridge Regression	20.2	19.5	7.1	3.4	124934
Lasso Regression	19.7	19.1	8.0	4.0	124965
Elastic Net	3.0	2.4	–2.5	–7.8	128440
Random Forest	86.0	85.2	11.1	8.9	–

Note: Adjusted R^2 values on the test set were, on average, weak across models, demonstrating the difficulty in predicting claim amounts from only structured data.

Baseline linear regression: It only accounted for 3.1% of the variation in claim amounts on the test set. Although the model accounted for almost 19.5% of the variance in the training data, the large discrepancy in performance on the test data suggests overfitting and poor generalizability. The high AIC value (141567) also indicates that the fitting result in relation to the model complexity is not good.

Ridge Regression: With Ridge Regression (i.e., with L2 regularisation), there was a slight generalisation improvement (Adj. R2 = 3.4%) and model complexity (as shown by the lower AIC (124,934)). It kept all features but reduced coefficients to prevent overfitting.

Lasso Regression: demonstrated the best test performance of the linear models (test Adj. R^2 = 4.0%). The L1 penalty on the other hand produced a more sparse and more interpretable model by shrinking some coefficients to zero. Lasso's marginal advantage over Ridge indicates that the automatic feature selection helped generalization.

Elastic Net: It was supposed to bring the best of both Lasso and Ridge, dramatically failed for simultaneous regression. It learned no useful patterns, as indicated by its negative R^2 scores on both the train and test sets (Negative scores may be caused by over-regularization). Its high AIC value also confirmed that the model was not optimal. These results indicate that hyperparameter tuning for Elastic Net is suboptimal, and that correcting it could lead to performance enhancement.

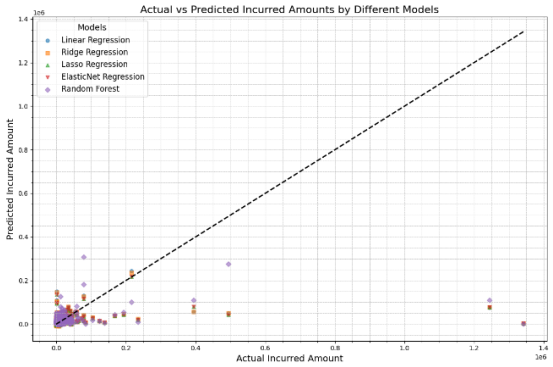
Random Forest Regression: It performed best, with a test Adjusted R^2 of 8.9%, just over double the test Adjusted R^2 of any linear model. Its high accuracy is attributed to the ability to capture the non-linear relationships and interactions between features. However, the discrepancy between training

(Adj. R^2 = 85.2%) and test (R^2 = 79%) performances suggests that curves are overfit. This notwithstanding, the model generalized better than the linear model and sensitively detected patterns that simpler models overlooked.

In conclusion, although Random Forest performed better than all other models, the best test Adjusted R^2 (8.9%) is modest. This is in line with the complexity of insurance claim data, where important predictive variables (i.e., detailed accident context or policy coverage), may be missed or not available. Furthermore, the tailed claims distribution is an obstacle for the model prediction accuracy - in particular large and rare claims with high costs are still systematically underestimated.

To assess how much our models learned from the data, we plotted actual vs predicted claim amounts (Figure 1).

Figure 1: Actual vs Predicted Incurred Amounts by Different Models on Test Set

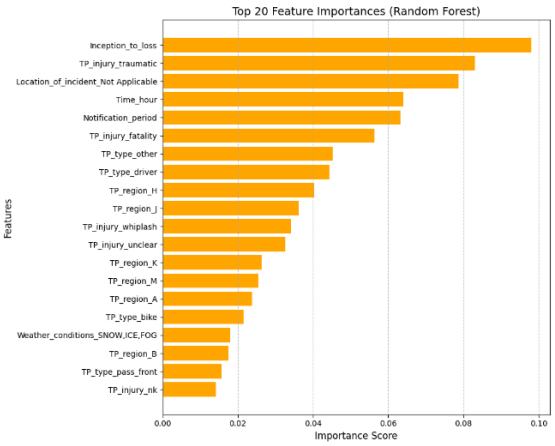


All the models had a large underprediction of high-cost claims, suggesting that they are not able to capture extremes (as also indicated by the low overall R^2). The linear models (Linear, Ridge, Lasso) made clustered predictions around the mean, ignoring the high variability of claims. Random Forest, on the other hand, had a wider prediction range and could better capture some of the higher claims, but it was still tending to underestimate the most extreme losses. This underscores that the Random Forest algorithm is better at modelling non-linear relationships, and at discriminating between moderate and high claims than at rare, high-cost events, the limitations of which are ongoing.

Feature Importance and Key Predictors

In order to visualize the Random Forest model and gain insight into its most important predictors, feature importance scores were examined along with SHAP (Shapley Additive Explanations) values (Figure 2). These techniques were used to determine which features contributed the most to the predicted insurance claim amounts at the global (across the dataset), as well as local (instance specific) levels.

Figure 2: Top 20 Feature Importances from the Random Forest model



The top predictor in the Random Forest model was **Policy Duration** (from Inception to Loss); longer durations could be associated with greater predicted claim amounts. SHAP analysis suggested that a longer policy period could be more challenging or with a higher degree of risk preserved. **Claim Reporting Delay** (Notification Period) was a powerful predictor as well (longer delays between incident and reporting tended to involve additional symptoms or severity, crowding in as a high cost or a quintile), and, for such high costs, a delayed reporting was more common.

The severity of the injury was a prevailing factor. (Binary indicators for fatal, traumatic, whiplash injuries were strong predictors. Fatal and trauma injuries were naturally associated with higher predicted costs because of the costs of associated medical and legal costs, and whiplash, despite being milder, occurred frequently enough to matter in the predictions. The **At-Fault Indicator** had an effect on the cost: for claims where the policyholder was at fault, cause a greater amount to be expected.

Some characteristics, such as **3rd party types** (e.g., pedestrians or cyclists), were not dominant at the global level but had fairly strong local effects, as a result of their vulnerability. Likewise, **Notifier type** (e.g. police) and **Vehicle motion status** were only important for local claims. On the whole, the model outputs matched reliable source scenario provider intuition, strengthening the credibility and interpretability of the results for practical purposes.

5. Conclusion and Discussion

This project focused on regression topic, formulating the problem of modelling the insurance claim amount prediction as a predictive modelling problem. The analysis indicated that the traditional linear regression techniques were interpretable but were not effective at modelling the claim severity, providing only 3–4% of the variance on novel data. Regularization methods, such as Ridge and Lasso, slightly enhanced performance by reducing the overfitting effect and introducing the

ability to select features. Yet the gains were slight. The Random Forest model had very significant test Adjusted R^2 (8.9%) which is almost double linear approaches. Its capacity to capture nonlinearity in relationships and interactions, in particular with injury severity, reporting delay and at-fault status, enabled it to more efficiently discriminate between high-and low-cost claims.

Although lead fraction correctly predicted the areas vulnerable to health impact relatively well, the prediction power in the global scale was low overall. The aspect of severe outliers and the skewed and heavy tail nature of the claim distribution did provide a significant challenge as all models had a tendency to underpredict the most expensive claims. Unsupervised analysis Limitation of accuracy: Incomplete datasets; Unhandled key predictive features Accident details (1,060 cases), Legal fact (1,060 cases), terms and conditions of the policy (660 cases); Insufficient models' explanation power.

The Random Forest model, however, provided many significant insights. Feature importance as well as SHAP analysis also showed that those who experience severe injuries, report late, and are at-fault policyholder ended up having a higher claim amount. These are well-known insurance properties, that increases the credibility and interpretability of the model, features that must be granted in practical context.

There were several limitations. The limited dataset (7,691 data instances) restricted the models for generalisation, especially the complex algorithms as well such as Random Forests. In addition, the analysis operated with default settings, or settings where only cursory tuning was applied, and Elastic Net performed poorly likely for being over-regularized. Future versions should have fine tuning, log transformation of the target and other modelling techniques which are specifically aimed at the upper tail to the distribution.

Augmenting the dataset with contextual and text labels, policy coverage specifics, and adopting more complex models (e.g. Gradient Boosting such as XGBoost) may greatly enhance the performance. Furthermore, the application of such models in insurance workflows should concentrate on early risks detection rather than accurate cost prediction with enhanced calibration to identify complex claims.

In conclusion, we show in this project that advanced machine learning techniques, particularly Random Forests, hold promise for claim severity, modelling provided richer data, focused methodologies, and fit to domain-specific use cases.

6. Appendices

Appendix A: Tables

Table A1. Summary Statistics of Incurred Claim Amount (after data cleaning)

Statistic	Value (in £)
Count of claims	7,691
Mean (average)	6,485
Median (50th percentile)	856
Standard Deviation	34,272
90th percentile	13,334
95th percentile	22,038
Maximum	1,341,914

Interpretation: The median claim cost of £856 indicates that half of all claims are below this value, while the much higher mean (~£6.5k) reflects the impact of a few exceptionally large claims. With 10% of claims exceeding £13.3k and a maximum of £1.34 million, the data is highly skewed, presenting significant challenges for predictive modelling due to the rarity and scale of extreme losses.

Table A2. Top Features by Importance in Random Forest Model

Feature (Predictor)	Importance Score (relative)
Inception_to_loss (days)	High ▲
Notification_period (days)	High ▲
TP_injury_fatality (fatal injury)	High ▲
TP_injury_traumatic (severe injury)	High ▲
PH_at_fault indicator (*insured at fault*)	Medium ▲
TP_injury_whiplash (whiplash injury)	Medium ▲
TP_type_pedestrian (pedestrian involved)	Medium ▲

TP_type_cyclist (cyclist involved)	Medium ▲
Location_of_incident (specific region code)	Low ▲
Vehicle_mobile (vehicle in motion)	Low ▲

Table summarises the highest-ranked predictors in terms of relative importance in the Random Forest model. Inception_to_loss, Notification_period, and fatal and traumatic injury severity of falls were the most positively influential features to claim costs. Fault status and third-party involvement (i.e. pedestrians and cyclists) made small contributions. Location and vehicle mobility along with other factors had lower overall global significance, but could play a role under certain conditions in prediction. Interpretation of rankings are provided based on SHAP.

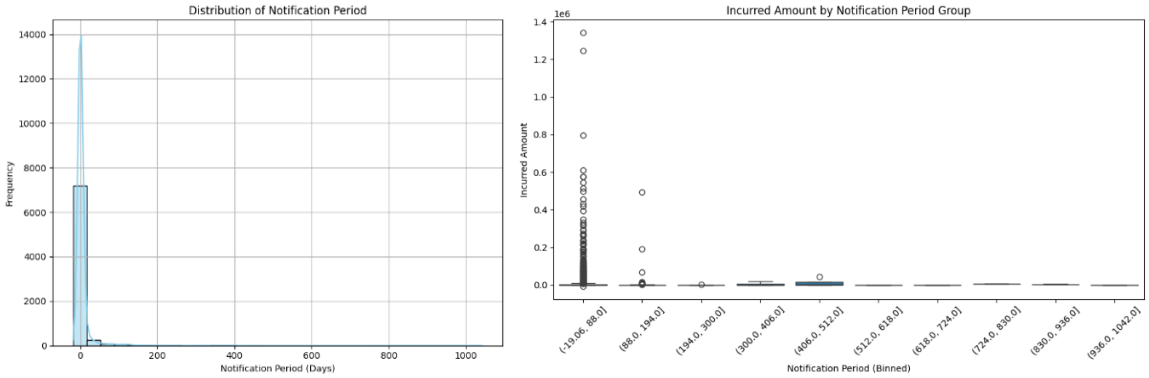
A3. Model Hyperparameters and Configuration

Model	Tuning Method	Key Hyperparameters	Notes
Linear Regression	None	–	Baseline model
Ridge Regression	Cross-validation (CV)	alpha = 50.0	L2 regularization
Lasso Regression	Cross-validation (CV)	alpha = 50.0, max_iter = 10000	Performs feature selection
ElasticNet	CV (Grid Search)	alpha = 100, l1_ratio = 1.0, max_iter = 10000	Behaved like Lasso in this case
Random Forest	Defaults used	n_estimators = 100, random_state = 42	Captures non-linearity, no tuning due to time

This table outlines the key configuration for each model used in the analysis. Linear Regression served as the untuned baseline. Regularised models (Ridge, Lasso, and ElasticNet) were tuned using cross-validation, with ElasticNet closely mirroring Lasso's behaviour. The Random Forest model employed default settings, balancing performance and computational efficiency within project constraints.

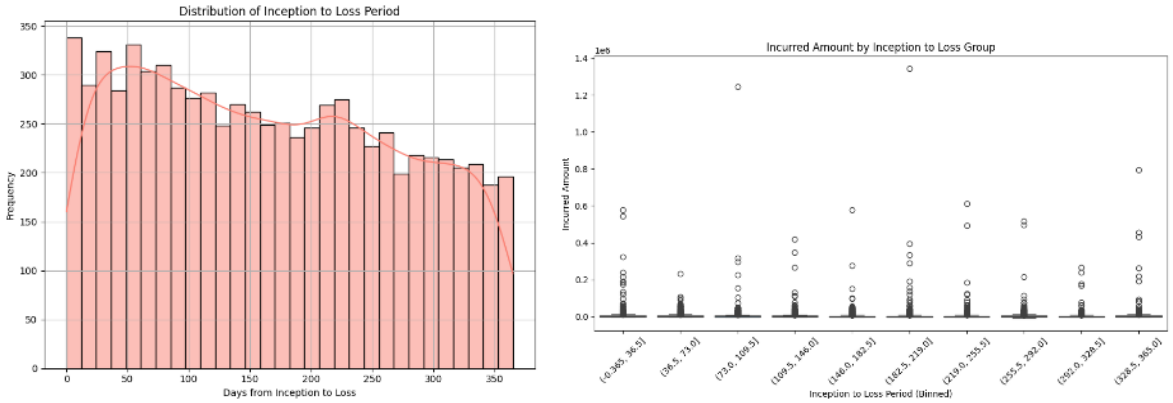
Appendix B: Graphs

Graph B1: Distribution and Incurred amount by Inception to loss period



Most claims are reported soon after the incident, as shown by the steep peak in the distribution. However, the second plot suggests that when claims are reported later, they can occasionally involve much higher costs, indicating that delayed notifications might be linked to more complex or severe cases.

B2. Distribution and Incurred Amount by notification period



First graph shows, most claims occur fairly evenly throughout the policy period, though there's a slight tendency for more incidents to happen earlier on. The second plot shows that high-cost claims can arise at any stage, regardless of how long the policy has been active. This suggests that the timing of a loss doesn't strongly determine claim severity, though rare extreme costs are present throughout.

7. References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232.

Henckaerts, R., Côté, M., Gianinazzi, T., & Yeo, J. (2018). A data science approach to modeling insurance losses using Gaussian processes. *Risks, 6*(3), 85.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, 4765–4774.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General), 135*(3), 370-384.

Staudt, Y., & Wagner, J. (2021). Assessing the performance of random forests for modeling claim severity in collision car insurance. *Risks, 9*(3), 53.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288.