Report for Statistical Machine Learning

Claim Severity Prediction Using ML Techniques

M.Sc Stochastics and Data Science

Department of Mathematics

December 16, 2022

1 INTRODUCTION

The report presents the performance of the different Machine Learning techniques that can be used for claim amount prediction in non-life insurance. It covers methods ranging from Linear Regression, Decision Trees up to Artificial Neural Networks. The framework of generalized additive models is extensively used in the insurance industry and therefore was taken as a standard against which other techniques were compared. The essence of the problem lies in the fact that insurance companies create reserves to meet future claims that may occur over the course of a particular financial year. So, if the reserves fall short of the claim amounts company can become insolvent. Hence, this problem is fundamental and significant to every insurance company and requires dedicated study.

1.1 Data

The data was taken from repository [1]. The data consists of policy-level information about the claims made by the policyholders for vehicle insurance products. There were missing values for a few variables which have been treated accordingly. The data set contains a total of 36176 claims and was divided into training and testing datasets. It was noticed that there were policyholders who made multiple claims, however for the simplicity of the models we have assumed independence. It was assumed that policyholders are covered for one year, which also generally is the time frame for vehicle insurance. The variables/predictors used to make claim amount predictions are certainly not sufficient, in practice, there are more variables like car type and size, location, fuel, use of vehicle, etc. On 21 December 2012, new EU rules entered into force, forbidding insurers to use a customer's sex as a premium determining factor. Accordingly, the policyholder's sex was dropped in order to make the model EU complaint. The data has been scaled for better analysis.

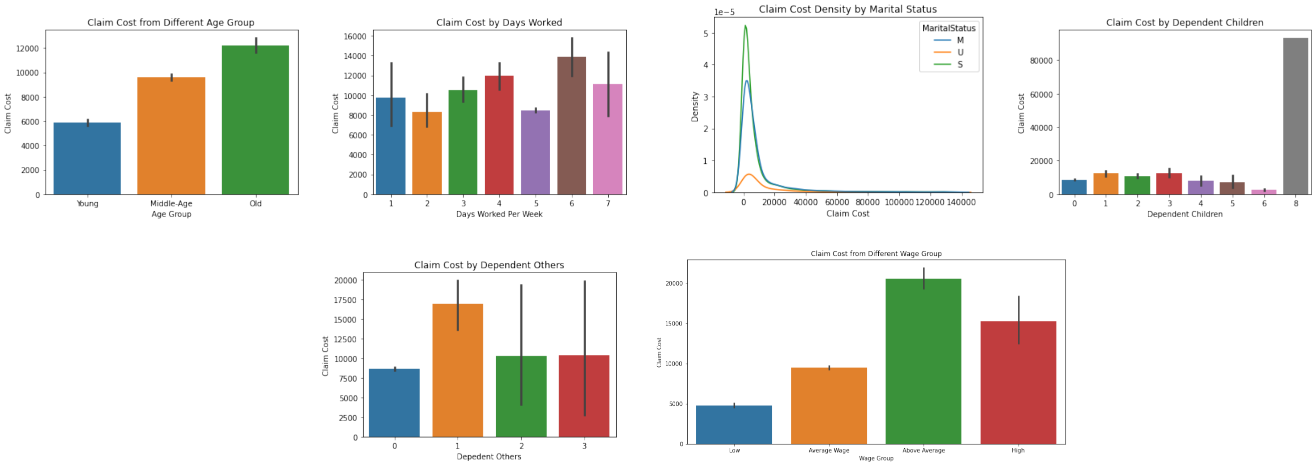
2 EXPLORATORY DATA ANALYSIS

The number of claims was observed to be lowest in 2006 and highest in 2001 i.e 0.41% and 5.7% of total claims respectively. Total claim cost was highest in 2005 and lowest in 2006. However, it was observed average claim amount was highest in 2006. Claims cost is positively skewed, which in general is observed in vehicle insurance since it is expected that there will be higher claims of low cost and lower claims of high cost. The variables "Dependents Children" and "Dependents Others" are highly skewed toward zero. There were high claims at an early age, which can be explained by the lack of driving experience. There exist some correlation between variables "Days Worked Per Week", and "Hours Worked per week". There exist multicollinearity between the variables "Days Worked Per Week", and "Hours Worked per week". It can be observed in the data we have some outliers, which might need some processing before claims settlement hence they have been removed from the data. Refer appendix for the correlation between the variables.

Table 1: Preliminary Analysis

Statistic	Age	Weekly Wages	Hours Worked Per Week	Days Worked Per Week	Initial Incurred Claims Cost	Ultimate Incurred Claim Cost
Count	35372	35372	35372	35372	35372	35372
Mean	33.79	412.66	37.27	4.89	7245.17	8727.68
S. Dev.	12.07	233.73	6.17	0.52	15616.29	16530.76
Min.	16.00	1.00	0.00	1.00	1.00	121.88
25 Percent	23.000000	200.00	38.00	5.00	662.50	912.04
50 Percent	32.000000	391.00	38.00	5.00	2000.00	3286.11
75 Percent	43.000000	500.00	40.00	5.00	9000.00	7913.09
Max.	79.000000	6453.00	640.00	7.00	690000.00	137241.12

Figure 1: Preliminary Visuals



[1] https://github.com/Sarah-2510/Vehicle-Insurance-Claim-Prediction

3 REGRESSION TECHNIQUES

3.1 Linear Regression

For the given training data the model was run for different combinations of predictors. However, to capture more information and comparison with other techniques full-scale model was fitted with all the variables. Below are the results of the final model.

Figure 2: Linear Regression Results

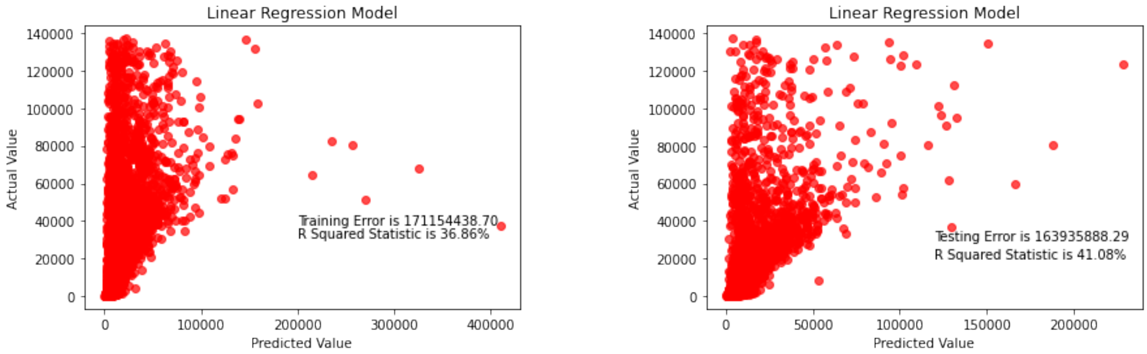
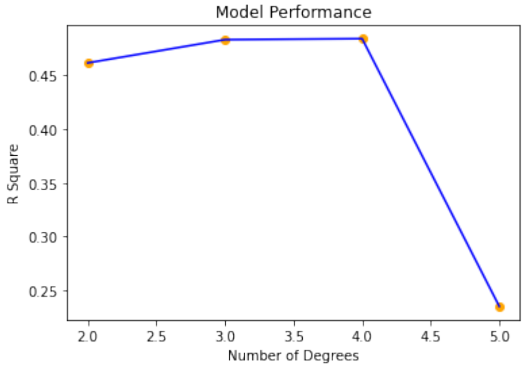


Figure 3: R Square Statistic

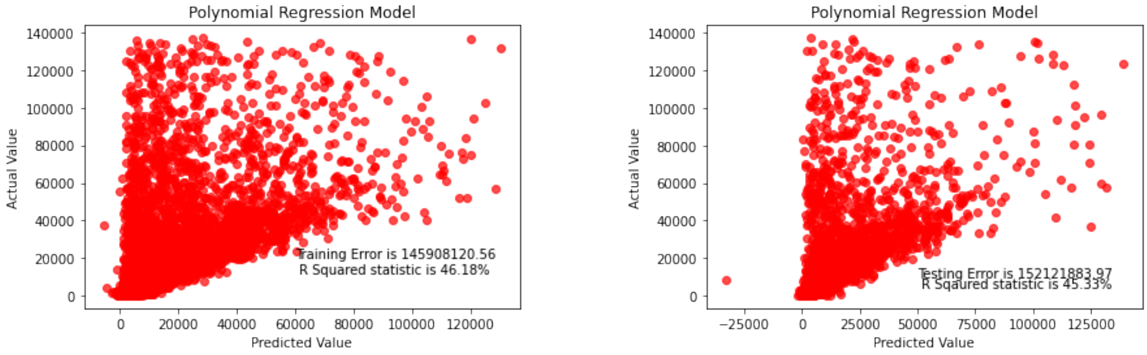


3.2 Beyond Linearity

3.2.1 Polynomial Regression

The method was applied for different degrees of polynomials for a full-scale model and in figure 2 is the R^2 statistic of the model as the number of degrees increases. Generally, the number of degrees is taken maximum up to 3 or 4 since a higher number of degrees can impact the shape of one region of data by far-off points and polynomial doesn't fit the threshold effect. In other words, the higher degree polynomial doesn't work very well at the tails. In fact, when a model with degrees more than 2 was fit to the testing data it doesn't work very well and there were way more negative predictions than the current model. The training and the testing error from the Polynomial model with degree 2 are given above. It can be seen that there is some improvement in R Square Statistic compared to linear regression.

Figure 4: Polynomial Regression Results



3.2.2 Generalized Additive Model

A natural way to extend the multiple linear regression model in order to allow for non-linear relationships between each feature and the response is to replace each linear component with a (smooth) non-linear function. Mathematically, it can be expressed as $y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + ... + f_p(x_{ip}) + \epsilon_i$. It is called an additive model because we calculate a separate f_j for each X_j, and then add together all of their contributions. Below are the results from the GAM model.

3.3 Tree-Based Methods

3.3.1 Bagging

In this method, we constructed B regression trees using B bootstrapped training sets, and average the resulting predictions. These trees are grown deep and are not pruned. Hence each individual tree has high variance, but low bias. Averaging these B trees reduces the variance. This technique improved the R^2 statistic significantly for the training data set.

Figure 5: Generalized Additive Model Results

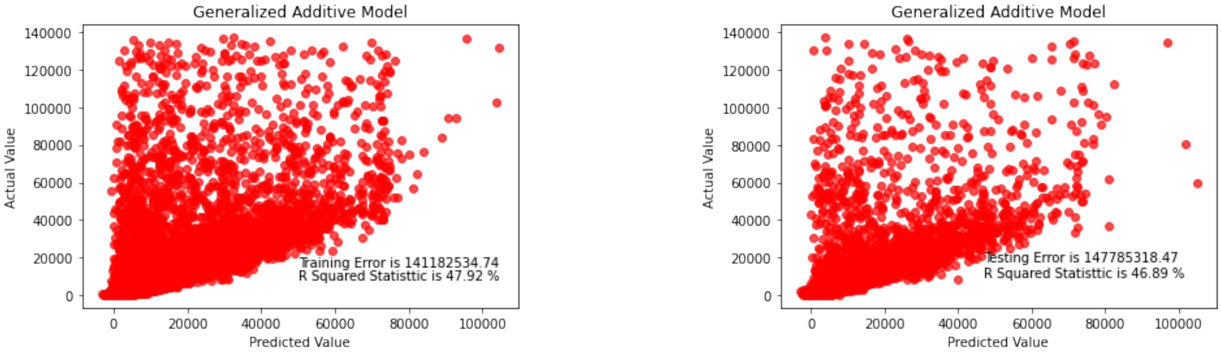
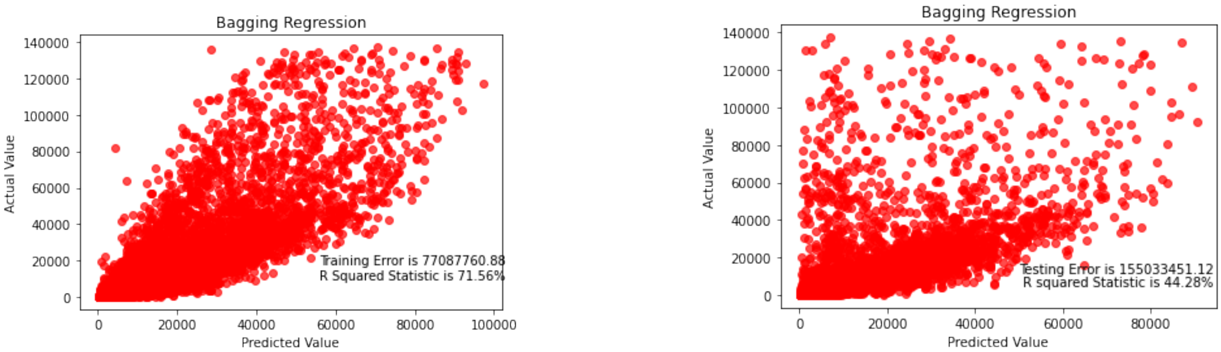


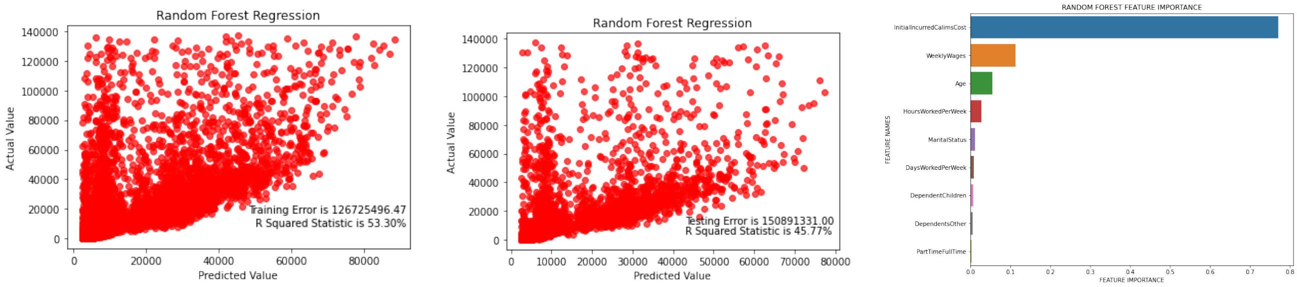
Figure 6: Bagging Regression Results



3.3.2 Random Forest

In this method, we build multiple trees but each tree is constructed with a subset of available predictors. This helps us to take into account the moderately strong or even weak predictors as a result trees are uncorrelated in comparison to bagging trees and hence improve the result. The main difference between bagging and random forests is the choice of predictor subset size m.

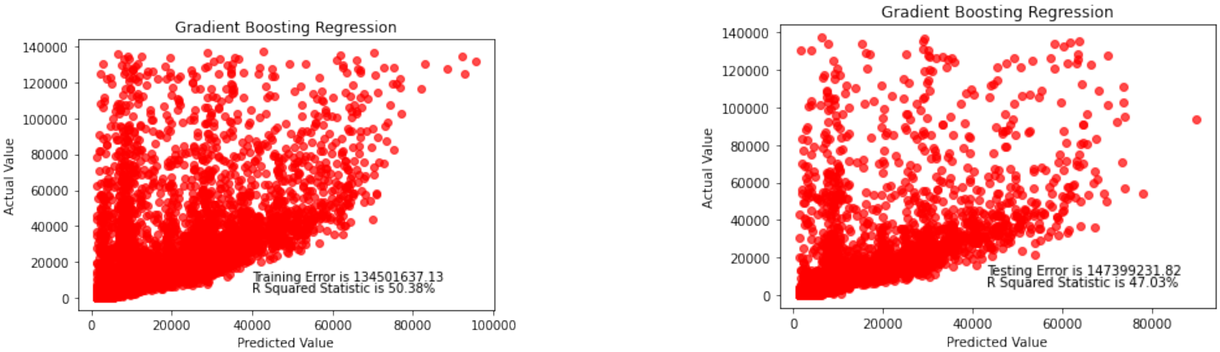
Figure 7: Random Forest Regression Results



3.3.3 Gradient Boosting

Boosting works in a similar way, except that the trees are grown sequentially: each tree is grown using information from previously grown trees. Boosting does not involve bootstrap sampling; instead, each tree is fit to a modified version of the original data set.

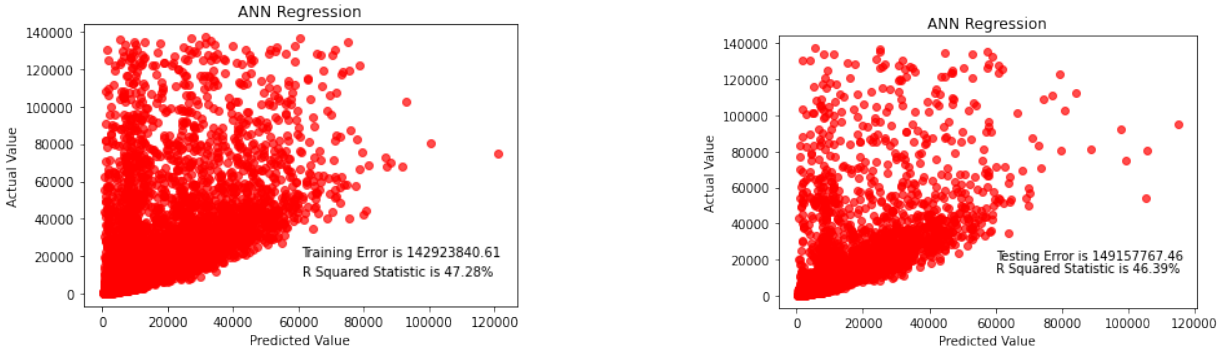
Figure 8: Gradient Boosting Regression Results



3.4 Artificial Neural Networks

For the given data we fit neural network model with two hidden layers and below are the results of ANN.

Figure 9: ANN Regression Results



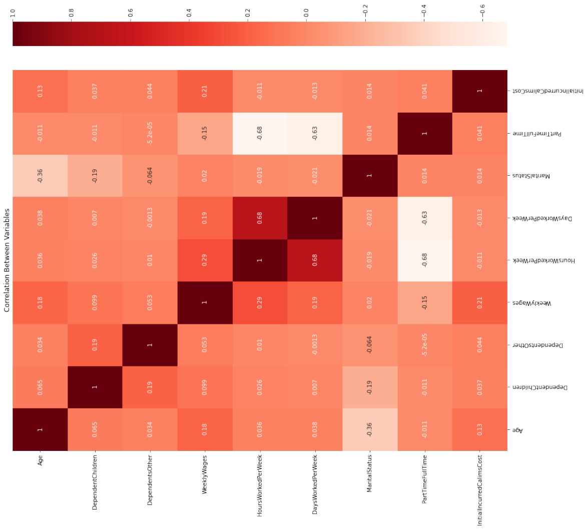
4 CONCLUSION

It can be observed that Bagging Regression has the lowest mean squared error for the training data set and somewhat similar statistics with other methods on testing data set. It was observed that when some outliers were included in the data set tree-based methods i.e Random forest and ANN performed best. It should be noted that we have included all the variables in our model, however, we can exclude some variables with high multicollinearity, this can also be noticed from the variable importance chart in random forest regression. We can say that given the complexity of the data tree-based methods and ANN can be used to model non-linear effects over traditional techniques like GAM. However, note that improvement in performance by using tree-based methods and ANN has a trade-off with the interpretability of the model. In our case, we can use GAM excluding variables with high multicollinearity.

5 APPENDIX

The correlation between the variables is given below

Figure 10: Correlation Between the Variables



5.1 Further Analysis

In the above Machine learning techniques, we did not remove the outliers. On taking the 90 percentile of values of claim cost we saw a significant change in the results. Below is the table comprising results from various methods.

Table 2: Results Without Outliers on Training Dataset

Techniques	MSE	RMSE	Coefficient of Determination (R^2 in %)
Linear Regression	6675003.51	2583.60	72.47
Polynomial Regression	6056452.90	2460.98	75.02
Generalized Additive Models	5532632.96	2352.15	77.18
Bagging Regressor	2981785.53	1726.78	87.7
Random Forest Regressor	5265211.33	2294.60	78.3
Gradient Boosting Regressor	5470295.12	2338.86	77.44
Artificial Neural Networks	5429909.17	2330.21	77.61

Table 3: Results Without Outliers on Test Dataset

Techniques	MSE	RMSE	Coefficient of Determination (R^2 in %)
Linear Regression	6651685.03	2579.08	73.69
Polynomial Regression	6059636.60	2461.63	76.03
Generalized Additive Models	5487990.09	2342.64	78.29
Bagging Regressor	5693908.04	2386.19	77.48
Random Forest Regressor	5710759.44	2389.71	77.41
Gradient Boosting Regressor	5656091.20	2378.25	77.63
Artificial Neural Networks	5478017.55	2340.51	78.33

References

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2021) *An Introduction to Statistical Learning, Second Addition*, Springer.

[2] Viktor Johanson (September 2021) *Tree-based methods for non-life insurance pricing*

[3] Sarah-2510 *Vehhicle-Insurance-Claim-Prediction*

[4] `https://github.com/Sarah-2510/Vehicle-Insurance-Claim-Prediction.git`