ISyE 6748 - Spring 2022
Practicum Project Report - Assurant Preventative

Team Member Name: Fanrui Yan

Project Title: Relationship Between Repair Cost and Vehicle

Contents

1 Problem Statement

Assurant Global Automotive develops, underwrites, markets and administers extended service contracts for global clients (dealers). This project focus on the U.S. market. Assurant has collected a lot of data regarding repairs of vehicles. I'm curious about the relationship between vehicle and the repair cost of them. This could be directly effect how to determine policy rates with clients. This process could help with understanding the cost of existing clients and future business.

In this project, I focus on vehicles 15 years or newer (2006 or later) since the older models would have been replaced as time goes and there are not much of them still running on the road now. I would pick a few relative features to analyze and to predict the total repair cost. I will use 4 different regression models to predict the cost, then use RMSE and R^2 to evaluate the performance.

2 The Data

2.1 Data Source

There are 7 datasets provided by Assurant for this project. They come from 2 sources, one of which is Global Warranty (GW) that contains all records from 2016-2021; 6 other datasets from GLOW that are also records from 2016-2021 (1 dataset for each year). There are 10,121,526 records in GW and 9,312,449 records in total in GLOW. The datasets contain slightly different features, and there are 33 features for the final combined dataset. Each row in the datasets is a part or service of a repair, there can be multiple rows for one repair. Variables in the data sets:

- Source: Global Warranty or GLOW
- Country: United States
- Status: Status of the claim
- Status_Group: Overall claims status
- Policy_Status: Policy status when exporting the data, not the status when claim happens
- Client_Name
- ClaimID: Unique for each claim, there are multiple rows per claim.
- PolicyID
- Currency: Currency of payment
- Claim_Start_Dt: Claim start date
- Date_In_Service
- Total_Unit_Count: Number of repairs in one claim
- Total_Authorized_Amount: Total authorized amount of a claim
- Total_Parts_Paid: Total amount paid on parts
- Part_Cost: Cost of a single part
- Part_Cost_Approved
- Total_Labor_Cost
- Inspection_Cost: This cost may or may not occur
- Total_Paid_Amount: Total repair cost, this is the response variable in this project
- Make_Desc: Make of the vehicle
- Model_Desc: Model of the vehicle
- Vehicle_Year: Year of the vehicle, there could be 2022 vehicles released in second half of 2021, thus with 2021 claim start date.

- New_Used: New, Near New, or Used
- Dealer_State: State where the policy was sold, not where vehicle was located
- Part_Description
- Component_Description
- Contract_Length: In year
- Tow_Paid
- Rental_Paid
- Paid_Date: Date when a payment towards the claim was made
- Repair_Facility_Name
- Total_Amount_Paid: Total amount paid for the claim
- Line_Item: What this row is: Inspection, Part or Labor. Only for GW

2.2 General Data Cleaning

My analysis focus on the relationship between total repair cost and the specs of vehicles, the following variables are not very helpful in my analysis, so I do not focus on them: Source, Status_Group, Policy_Status, Client_Name, PolicyID, Claim_Start_Dt, Date_In_Service, Dealer_State, Paid_Date, Repair_Facility_Name, Tow_Paid, Rental_Paid. For example, the date column for GLOW 2016 is missing date information, per Alejandra, we can leave it out, but since I'm not using date variables I will keep the dataset in my project so I have more data for modeling. The part related information, though very valuable, are not what I am analyzing here because my analysis focus on total repair cost. This leads me potentially to use the following predictors for modeling: Total_Paid_Amount, Make_Desc, Model_Desc, Vehicle_Year, and New_Used. Other variables are also used for data cleaning purpose.

Currency
The data should be for the U.S. only, I checked the Country column that all data are showing US. But there are some data showing other currency in the Currency column, since I'm analysing the total repair cost, I would need an accurate amount on the cost, different currency would make it hard to justify the true cost of repair (the exchange rate changes all the time so a simple conversion to USD would not work). Thus I have removed data that are not "USD" or "US Dollar" in the Currency column. The total records removed is 4749 (0.024% of the data).

Location
According to the project guideline, we should not consider any repair that's been done at Puerto Rico, so I removed records with "San Juan", "Puerto Rico", and "Bella" in the Repair Facility column. There are 328,907 records removed (1.692% of the data).

Shifted Columns
Some of the data have the last few columns shifted, this caused information for Repair Facility to be lost, and we would not know if the repairs have been done at Puerto Rico or not, so I have to drop these records. The data with shifted columns would have string in Contract Length column, so I dropped these 273 records (0.0014% of the data).

Contract Length
There are 128 records (0.00066% of the data) with negative contract length, these records should be removed. There are also records with very long contract length, those should be for the unlimited warranty policies, which are not our concern, so I will keep them.

Status of the claims
There are 21 different status in the data set, some are the same status but used different wording. But there

are 51,353 data (0.264% of the data) in the following status that affect payment, thus should be dropped from the analysis:

- Cancelled: cancelled claim, no payment made
- Void: voided records, no payment made
- Sent Back for More Info: send back for more info, no payment should be made at the time
- Sendback: same status as above just different wording
- Correction Needed: while waiting on correction, there should not be payment made
- Other: a handful on this status, drop to avoid confusion.

Non-Positive Total paid amount
There are 448,090 records (2.3% of the data) with negative or 0 in the total paid amount column. The negative ones should be removed per conversation with Assurant. I will take the total repair cost per claim ID in the next step, and data records with 0 amount will be added into the total cost according to its claim ID. But this takes longer time to compute, so I decide to drop them here to save some memory and time later.

The total number of data dropped up to this step is 833,500, which is 4.29% of the data.

Total Cost per Claim ID
Now, I will total the repair cost by Claim IDs. GW and GLOW are treated separately: for GW the total cost per claim ID is just the total of "Total_Paid_Amount" of each claim ID. But for GLOW, the "Total_Paid_Amount" is cumulative amount, so I need to take the maximum "Total_Paid_Amount" and divide by the "Total_Unit_Count" to get the total cost of each claim ID. Now I have 3,539,017 claims each with its total repair cost and the basic vehicle information - make, model, year, and condition.

2.3 Detailed Cleaning

Outliers
There is a stand out claim which shows a total repair cost of few million dollars, with review of the claim information, it must be a typo and this data should be removed. Regarding other outliers, due to the large size of the data set, I would not try to plot boxplot or QQ-plot. Instead, I use the standard $mean \pm 6 * std$ to determine outliers. There are 1361 outliers (0.038% of the data) that need to be removed.

Age of Vehicles
Consider the ever changing market and the technology innovation, I would focus my analysis on the vehicles less than 15 years old (2006). Now I have 2,897,375 claims in total.

Text Cleaning
Since the data come from 2 different sources, there are some discrepancies in the wording and I would like the wording to be matched. First, I changed all the words to upper case. Then I reviewed and updated "New_Used" column by changing "Near New" to "Near_New". Next, I checked the "Make_Desc" column and there are over 100 categories, I could manually review and update the data. Some wording are different ways of writing the brand like "BMW" and "B.M.W."; some are brand and vehicle model combined and listed in the Make_Desc column, I split those and write the vehicle model information to the Model_Desc column. There are 6,000+ different vehicle models in the "Model_Desc" column, I couldn't complete the data cleaning step of this predictor due to the time limitation.

Missing Data
There are 19 claims with empty vehicle make, and 4,045 with "null" or "Unknown Make" in the vehicle make column, I have to drop these data since they do not provide any useful information. There are 142 claims with "null" in the "New_Used" column which should be removed as well. The total removed data from this step is 4206 (0.145%).

Year
The 2023 and future vehicles should not have been released to the market yet as of 2021, so any claim with "Vehicle_Year" later than 2022 should be removed as well. The total claims found with this issue is 86 claims (0.00297%). My final data set contains 2,893,077 claims.

2.4 Exploratory Data Analysis

Outliers
Look into the outliers a bit further, the brand of Land Rover stands out as it counted over 2 times more than the second highest count by vehicle make. Not surprisingly, used cars also cost more in the high cost end. It is interesting that the year 2016 has much more high cost repairs than other years, we may want to dive deeper what has happened in 2016. Most of the outliers are engine repairs.

Table 1: Count by vehicle makes

Make	Land Rover	Chevrolet	Nissan	Ford	Jaguar
Number	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Table 2: Count by New_Used

New_Used	Used	New	Near New
Number	[redacted]	[redacted]	[redacted]

Table 3: Count by year

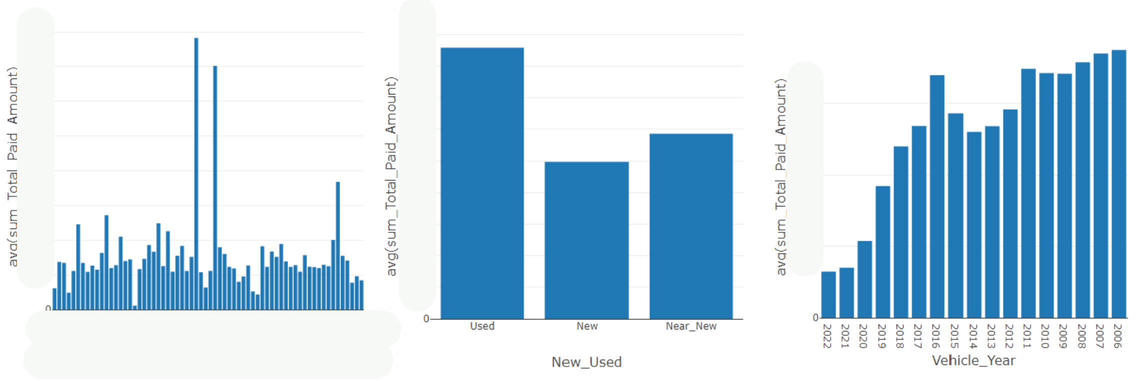
Year	2016	2011	2015	2012	2013
Number	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Summary Statistics
From the summary statistics below, I could conclude, even though some of the repair cost could get pretty high, the majority repair cost is within a reasonable range.

Table 4: Summary Statistics for the response variable "Total_Paid_Amount"

Percentile	25th	Median	75th	90th	95th
Number	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Mean by Predictor
We can see from plots (a) and (b) in Figure 1, that the means for vehicle make and vehicle condition (New_Used) are different for each category, so these 2 variables should have strong predictive power to the total repair cost. And there seems to be an increasing relationship between vehicle age and repair cost from plot (c) in Figure 1. Since the "Model_Desc" variable has not been cleaned, I would not include that in the model. An interesting finding is that the 2 tall bars plot (a) are Land Rover and Jaguar, we could see the average total repair cost for them are much higher than any other brand.



(a) Cost vs. Make

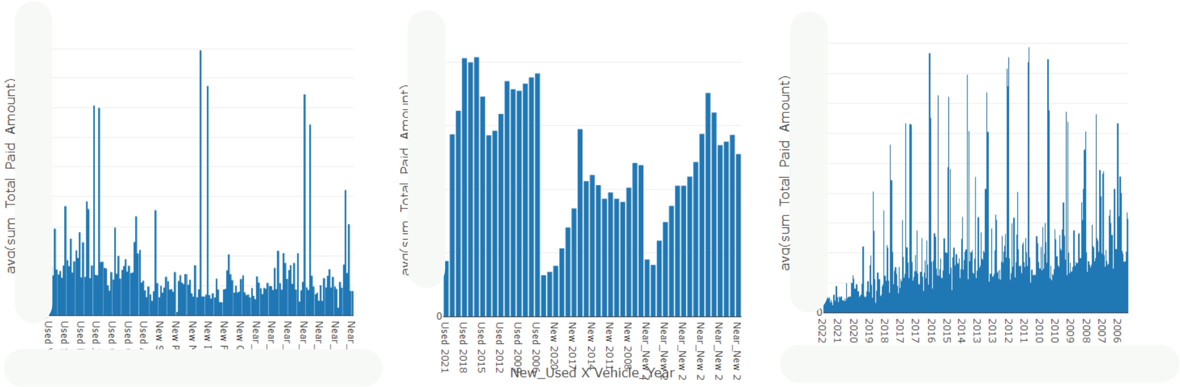
(b) Cost vs. New_Used

(c) Cost vs. Year

Figure 1: Average Total Repair Cost

I'm also interested in the interaction effects of the predictors. From plot (a) in Figure 2, we could see, on average, used vehicles have higher repair cost and new vehicles have the least amount of repair cost. We could also see that there are 2 very tall bars in each condition, those 2 brands are also Land Rover and Jaguar. Similarly, on plot (c) in Figure 2, we could find 2 tallest bars each year, and they are again Land Rover and Jaguar. So these two brands are expensive to repair regardless of the vehicle condition and age.

Another interesting finding is in plot (b), we could see there are 3 sections, one for each vehicle condition. There are increasing relationships for age and repair cost for New and Near new vehicles. However, this relationship is not obvious for Used vehicles. Despite the less than 1 year old 2021 used vehicles, the average repair cost for used vehicles went up high starting on age 2.



(a) New_Used X Make

(b) New_Used X Vehicle_Year

(c) Year X Make

Figure 2: Average Total Repair Cost by Interaction Term

Multicollinearity

I calculated the VIF for the original predictors only and interaction terms included. There isn't any multicollinearity issue with the original predictors. But there are multicollinearity issues for "New_Used X Make_Desc" and "Make_Desc" features. Since high correlation among interaction terms and main effects is normal and expected, the multicollinearity test is acceptable.

Table 5: VIF

Predictor	VIF 1	VIF 2
Vehicle_Year	2.398748	3.552076
New_Used	1.509696	6.709537
Make_Desc	1.904268	17.548861
New_Used X Vehicle_Year	-	6.545062
New_Used X Make_Desc	-	17.403177
Vehicle_Year X Make_Desc	-	8.154296

3 Methodology

Since the response variable for this project is continuous number, I would use several regression models to predict the repair cost - "Total_Paid_Amount". I used StringIndexer to map the string columns to label indices. Then I randomly split the data set into 2 sets, a training set that contains 85% (2,459,115) of the data and the rest to be in the testing set (433,962). I would train the models using training set and test the models using testing set.

3.1 Ordinary Linear Regression

Ordinary Linear Regression has good predictive ability, it could give the relative influence of predictors. However, it won't perform well if predictors are not in linear relationship with response variable.

3.2 Ridge Regression

Ridge Regression uses L2 penalty term which could reduce the effect of multicollinearity for the model with interation terms. Using 10 fold cross validation, the result training RMSE are 2561.18 for original features only, and 2552.17 for model with interaction terms included.

3.3 XGBoost

Extreme Gradient Boosting is an efficient implementation of gradient boosting that works great in regression predictive modeling. I wanted to perform grid search to find the best hyperparameters, but it takes too much of time and the cloud would turn off after a while. So I tested a few values and picked the best out of them. Using repeated stratified 10-fold cross-validation with three repeats to train the model and resulted training RMSE of 2334.87.

3.4 Random Forest Regression

Random Forest Regression works well for large data set, it reduces overfitting and improve accuracy. A downside would be that Random Forest Regression would not be able to predict values that are not within the training set range. Similar to XgBoosting, I could not perform grid search for this model, I had to manually pick and choose the hyperparamenters. Using repeated stratified 10-fold cross-validation with three repeats to train the model and resulted training RMSE of 2333.31.

4 Results & Evaluation

Using RMSE and R-squared for testing set to evaluate the models, Random Forest Regression with original features and interaction terms perform the best out of 4 models. It was in nature that the more feature added to the model the better the prediction is. I wouldn't say that there are significant difference in predicting power of the models with or without interaction terms. Even though Random Forest Regression performs the best, but it has it's own limitations. I believe the predicting power can be improved if more vehicle information could be added to the model. For example, A cleaned Model_Desc feature, millage, and color of vehicles.

Table 6: Comparison of Model Performance

	Model	RMSE	Rˆ2
Original	Linear Regression	2563.02	0.0206
	Ridge Regression	2563.58	0.021
	XgBoost Regression	2338.55	0.1847
	Random Forest Regression	2337.72	0.1853
With Interaction	Linear Regression	2552.74	0.027
	Ridge Regression	2552.74	0.0283
	XgBoost Regression	2333.53	0.1877
	Random Forest Regression	2332.68	0.1882

5 Challenges

This real world project is very challenging, I have some valuable findings in the analysis process, although the final model is not a perfect model that matches all the assumptions and gives accurate prediction. The greatest challenge for this project is time and business knowledge. I spent about 2 months to get to know the data/variables from business perspective and I have to revise my work accordingly each day. A more efficient way of communication would be helpful to speed up this step. And by the time when I would be working on the cleaning of ”Make_Desc” features, I noticed I would not have enough time to complete the cleaning since there are 6000+ different vehicle models, so I have to left it out from my model. Another challenge is a lack of data, even though there are 30+ variables but only a few of them are useful for my model. It would be helpful if more information like millage can be included, just like what Yingying has said. Last but not least challenge is the technical challenge, the cloud that we are running our script on has limited memory and it is quite slow. When I run things like GridSearch or Cross Validation, sometime it would time out due to speed and memory. As a result, I have to limit my computation, for example, I manually test other than use grid search to find the best hyperparameter values.

6 Findings and Suggestions

I found that Rand Rover and Jaguar on average cost the most to repair. Vehicles with year 2016 seems to cost more in repairing. Used vehicles cost a lot more to repair than new and near new vehicles regardless of age. One should consider these factors when determining policy rates for related vehicles/clients.

A Models

A.1 Ordinary Linear Regression

A.1.1 Without Interaction

Coefficients	Value
Interception	46975.86
Vehicle_Year	-22.62
New_Used	-395.74
Make_Desc	31.72

A.1.2 With Interaction

Coefficients	Value
Interception	41324.61
Vehicle_Year	-19.83
New_Used	-807.47
Make_Desc	56.77
New_Used X Vehicle_Year	29
New_Used X Make_Desc	10.13
Vehicle_Year X Make_Desc	-4.59

A.2 Ridge Regression

A.2.1 Without Interaction

Coefficients	Value
Interception	46976.29
Vehicle_Year	-22.62
New_Used	-395.74
Make_Desc	31.72

A.2.2 With Interaction

Coefficients	Value
Interception	41326.72
Vehicle_Year	-19.83
New_Used	-807.44
Make_Desc	56.76
New_Used X Vehicle_Year	29
New_Used X Make_Desc	10.13
Vehicle_Year X Make_Desc	-4.59