AGGREGATE LOSS PREDICTION USING MULTIPLE-CLASS CLASSIFICATION TECHNIQUES

A Thesis

Presented to the Faculty of the Department of Mathematical Sciences

Middle Tennessee State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Mathematical Sciences

by

Chuanlong Zhang

June 30 2021

Thesis Committee:

Don Hong, Chair

Vajira Manathunga

Qiang Wu

Lu Xiong

APPROVAL

This is to certify that the Graduate Committee of

Chuanlong Zhang

met on the

30th day of June, 2021.

The committee read and examined his thesis, supervised his defense of it in an oral examination, and decided to recommend that his study should be submitted to the Graduate Council, in partial fulfillment of the requirements for the degree of Master of Science in Mathematics.

Dr. Don Hong
Chair, Graduate Committee

Dr. Vajira Manathunga, Committee Member

Dr. Qiang Wu, Committee Member

Dr. Lu Xiong, Committee Member

Dr. James Hart
Graduate Coordinator,
MS-Math Program

Dr. David Chris Stephens
Chair, Department of Mathematical Sciences

Signed on behalf of
the Graduate Council

Dr. David Butler
Dean, School of Graduate Studies

DEDICATION

I would like to dedicate this thesis to my parents for their unconditional love.

ACKNOWLEDGMENTS

I appreciate my advisor and thesis committee for the unremitting efforts and help in the process of the accomplishment of my thesis and kind understanding and support during the winter 2020 and spring 2021 when I experienced a serious health issue. I appreciate all professors in the Department of Mathematical Sciences who taught me or gave me suggestions and guidance during my studies at Middle Tennessee State University. I am also grateful to all fellows who help enhance my programming skills. Finally I appreciate the rich and accessible academic resources offered by Middle Tennessee State University.

ABSTRACT

In this thesis, we consider a model for predicting future losses in car insurance applications by using multiple-class classification algorithms.

In the car insurance payment data, the records can be divided into four different groups based on the different types of payments, such as labor-dominant (LD), parts-dominant (PD), other-dominant (OD), and none-dominant (ND) payments. We first apply the multi-class classification algorithm to predict the probabilities that a randomly selected subject belongs to the corresponding group, then, for simplicity, the predicted payment amount can be calculated by using the total expectation formula after determining the conditional expected payment amount using training data. This method is easy to implement and yield satisfactory prediction accuracy.

We compared proposed model accuracy against general linear regression and classical individual aggregate loss models accuracy for the test dataset. The comparison results show that the proposed model outperforms other models in accuracy for given test datasets. The matrix used to describe the distribution of the true groups and the distribution of the predicted groups indicates there exists a relationship between the payment amount and the type of group it belongs to.

CONTENTS

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

1.1 Background

The dataset we use in this study is based on a moving truck under a basic warranty. The dataset consists of information including truck's identity information, number of claims, each claim's loss amount (usually the insurance payment), among others. During the insured period, basic truck warranty covers most losses of a new truck during a certain period of time since the truck is sold.

Our purpose is to make a better prediction for the future loss of the truck so that rate making for insurance policies become more efficient and accurate. Since there are several features of a basic truck warranty policy which make it different from a general auto insurance policy, different techniques can be applied to its ratemaking other than traditional method such as loss triangle method [1].

First, compared to the maintenance cost, liability is a significant portion of payment of general auto insurance policy due to the negligence of the driver. Instead, the maintenance cost is the most significant portion of a basic truck warranty, since it covers engine, systems and parts of the insured truck with a few exceptions such as routine oil change.

Second, factors to determine the premium of a general auto insurance policy are involved, such as the age of the driver; the driving history of the driver (having traffic accident(s) or not); the year of the insured auto; the mileage of the insured auto; etc. All these factors will cause the difference in premiums from one general auto insurance policy to the other. In essence, general auto insurance premium depends on the driver and vehicle both. This discrimination is fair and allowed in the insurance industry. However, the premium of a basic truck warranty is basically the same. In fact, sometimes there is no extra charge for the basic truck warranty since the premium has been incorporated in the cost to purchase the truck. Therefore, identifying the risk stemming from the driver is not as important as general auto insurance. Therefore, figuring out the average loss amount is meaningful enough.

Besides two features mentioned above, basic truck warranty policy is also different from the extended truck warranty policy. Extended warranty policies always come into effect after the basic warranty policies of insured trucks expire. There will be extra charge for extended warranty policies and the charge might be different based on the usage condition of the truck.

1.2 Payment Portion

Each payment of a truck warranty policy can be divided into three portions: labor payment; parts payment; and other payment.

In general, each payment could be affected by the deductible and limit of their policy. Besides these two factors, there are also some other factors affecting the payment.

Labor payment is always incurred by the mechanics when repairing the truck. The amount highly depends on the number of hours mechanics took to repair the truck. In general, longer repairing time always indicates a higher labor payment. Labor payment is also affected by other factors such as garage shop.

Parts payment is always incurred when one or more parts are replaced. The amount highly depends on the price of new parts. Some researchers would like to analyze property and casualty topics by survival analysis techniques [2, 3]. We can have the similar thought for parts payment since each part has a distribution for its life time.

Other payment is more miscellaneous than labor payment and parts payment. One example of other payment could be a towing fee from a road to a garage due to the disfunction of the truck while it is driven. Since the cause for this portion of payment varies, it is hard to draw some conclusions about the factors that can affect this portion.

1.3 Multiple-Class Classification

In statistics, a classification problem identifies to which group one or more observations belong, given that there are two or more groups, and each observation is from exactly one of these groups. In statistics, the terminology of a so-called group is class [4].

One typical example of a classification problem is to identify if an email is a spam or not. There are only two classes: Spam and Non-spam. This is called binary classification. When the number of classes is more than two, the classification is a multiclass classification, or multiple-class classification.

There are several algorithms to perform classification. In general, each algorithm requires a dataset (named training set) to the model. Then, the model can perform predictions for other similar data sets.

The outputs of some algorithms, such as logistic regression, are probabilities that each observation belongs to different classes, then, a threshold probability (such as 50%) will be applied to identify each observation into classes. However, the outputs of some other algorithms, such as linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) is only the class each observation belongs to. In our model, multiple-class logistic regression will be implemented.

The foundation of logistic regression/classification model is derived from multiple linear regression. For binary logistic classification, the logarithm of odds is predicted by multiple linear regression model. Suppose we have p predictors, $\mathbf{X} = (X_1, X_2, ..., X_p)$, the logarithm of odd, which is $\log(\frac{p}{1-p})$ is predicted by the following formula:

$$\log(\tfrac{p}{1-p}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_p X_p$$

In other words, the probability is predicted by

$$p = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_p X_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_p X_p}}$$

For multinomial logistic classification, suppose there are n classes, this means n probabilities can be predicted. Then we always define a reference class (take class 1 for example), the logarithm of the ratio of probability of class i over the probability of class 1 is

$$\log(\frac{p_i}{p_1}) = \hat{\beta}_{0i} + \hat{\beta}_{1i} X_1 + ... + \hat{\beta}_{pi} X_p = \mathbf{X}^T \hat{\boldsymbol{\beta}}_i \text{ for } i = 2, ..., n$$

where $\hat{\boldsymbol{\beta}}_i$ is the estimated coefficient vector for class i.
Eventually, the estimated probabilities are expressed as

$$\hat{p}_1 = \frac{1}{1 + \sum\limits_{i=2}^{n} exp(\mathbf{X}^T \hat{\boldsymbol{\beta}}_i)} \text{ for class 1}$$

and

$$\hat{p}_i = \frac{exp(\mathbf{X}^T \hat{\boldsymbol{\beta}}_i)}{1 + \sum\limits_{i=2}^{n} exp(\mathbf{X}^T \hat{\boldsymbol{\beta}}_i)} \text{ for } i = 2, ..., n.$$

1.4 Framework of Following Chapters

In Chapter 2, we are going to introduce the data and preprocess the data for the project. In Chapter 3, we will propose our model and make comparisons with other two models. In Chapter 4, we will draw some conclusions based on the results from Chapter 3 and propose some potential improvements and future works.

CHAPTER 2

DATA AND DATA EXPLORATION

2.1 Introduction to The Data

The dataset we are going to analyze is simulated by a real dataset. The real dataset contains more than 70 columns. But some columns contain descriptive information rather than numerical or categorical data, and some columns contain a significant percentage of missing data, which also cannot be taken advantage as fully as we can. Eventually, we only selected 14 columns for simulation and do further data analysis. Table 1 shows the columns (variables) we simulated.

11766 records were simulated from the real data. Since there exists recording mistakes, the simulated data has some problematic records, and they should be corrected before the further analysis.

2.2 Data Preprocessing

First of all, we correct the data.

172 records have earlier FAILDAT than the REPADAT. This means the insurer had covered the loss before these trucks got repaired, which is impossible. 3 records have FAILDAT 2 years after the trucks were sold. These records should be categoried in other warranty policies (such as extended warranty policy) instead of basic warranty policy. 1 record has 0 total payment, which means these claims are not covered by the warranty policy. These records should be removed since we only focus on the claims with non-zero payments. We believe these records contain obvious recording mistakes and do occupy an insignificant proportion of the entire simulated data. So we could remove these 176 records. After these 176 records were removed, there are 11590 records. Next, we are going to add and remove some variables based on the

Table 1: Description of Simulated Variables

Variable Name	Description
DEALERCOD	Categorical variable with 474 categories, numeric description of each dealer
SALESMODL	Categorical variable with 43 categories, alpha-numeric description of each model of truck
CHANNEL	Categorical variable with 3 categories, alpha-numeric description of each brand of truck
SERIAL.	Unique alpha-numeric description of each truck
WARTYPE	Categorical variable with 3 categories, numeric description of warranty type
DELDAT	The date when the truck was sold
FAILDAT	The date when the truck got repaired
REPADAT	The date when the cost was transacted from the insurer to the garage
DEFECTCOD	Categorical variable with 73 categories, numeric description of the reason for the issue of the truck
HOURMETER	Numeric variable, numeric description of number of miles the truck has been driven until the truck was repaired
LABORPD	Numeric variable, labor payment
PARTSPD	Numeric variable, parts payment
OTHERPD	Numeric variable, other payment
TOTALPD	Numeric variable, total payment

given data.

First, we will remove DEALERCOD. Note the DEALERCOD is a variable with 474 categories. We expected the model we are going to construct can indicate some dealers which sell trucks with fewer future payment amounts (or more future loss amount) before we simulate the data. In fact, there are too many dealers. On average, there are only 24.5 records for each categories. This variable could get better usage if there were more records (such as 1000 records per dealer). So, this variable would be removed.

Second, we will retain SALESMODL and remove CHANNEL. Each brand contains several models, but each model can only come from one of these three models.

So, we do not need both of these variables. Since SALESMODL contains more specific information than CHANNEL, we prefer to retain SALESMODL and remove CHANNEL.

Third, we will remove WARTYPE. Even though there are three categories, two of these categories only contain one record and five records only (note that there are 11590 records so far, and there are 15087 trucks with no payment during the warranty period). And the amount of total payments of these six records are not important (too high or too low relative to the average level of total payment).

Fourth, we will create a new variable to count the days between when the truck was sold and when the truck was repaired. Since the insured trucks by this warranty policy are all new trucks, we could treat DELDAT as their "birthday" (the first day they were able to be driven). The number of days between DELDAT and FAILDAT might indicate the distribution(s) of the life span(s) of some parts in the truck. We call the variable we created is deal.to.fail.

Fifth, we will remove REPADAT. In general, larger claims need more time to be processed. So, the number of days between DELDAT and REPADAT could be an indicator of the size of payment amount. However, we have to discard this potential indicator. The reasons are as follows: first, this variable is highly similar with the newly-created variable we mentioned in the last paragraph (the number of days between DELDAT and FAILDAT), since the correlation coefficient is 0.9984191, which causes a multicollinearity problem. Second, in the real world, we cannot exactly know the value of this variable until the insurer pays to the garage because we do not know REPADAT.

Sixth, we need three variables to indicate the ratios of labor/ parts/ other payment over the total payment. Three variables were created by dividing LABORPD/ PARTSPD/ OTHERPD over TOTALPD. Based on these three variables, we could divide all observations into four groups. The payments in these four groups are labor-dominated, parts-dominated, other-dominated, and none of them are dominated. For our convenience, we will call them as labor-dominant group (LD), parts-dominant group (PD), other-dominant group (OD), and none-dominant group (ND), respectively. The next question is how to set up the criteria to divide the data into the four groups. After several attempts by the model we are going to introduce, we select 75% as our criteria, i.e. if the ratio of LABORPD over TOTALPD is larger than 75%, the observation will be categorized into the labor-dominant group (LD); if the ratio of PARTSPD over TOTALPD is larger than 75%, the observation will be categorized into the parts-dominant group (PD); if the ratio of OTHERPD over TOTALPD is larger than 75%, the observation will be categorized into the other-dominant group (OD); if none of these three ratios is larger than 75%, the observation will be categorized into the none-dominant group (ND).

In addition, it is worth-noting that there are 6 unavailable variables when we construct the predictive model to predict TOTALPD. They are LABORPD, PARTSPD, OTHERPD, and the ratios of them over TOTALPD (the three new-created variable we mentioned in the last paragraph). Because we can never know the ratios until we know the value of TOTALPD. And knowing LABORPD, PARTSPD and OTHERPD and predicting TOTALPD does not make sense.

So far, the only four variables we can take advantage to predict the total payment are SALESMODL, deal.to.fail, DEFECTCOD and HOURMETER. Two of them (SALESMODL and DEFECTCOD) are high-dimension variables (variables with a decent amount of categories). This makes constructing the predictive model more challenging.

2.3 Exploratory Analysis

2.3.1 Variable 1: group.number

Using R and Excel, we could get more information about these selected variables. The following table shows the summary information about the TOTALPD from the entire dataset, labor-dominant (LD), parts-dominant (PD), other-dominant (OD), and none-dominant (ND) groups. It is shown in Table 2.

Table 2: Quantiles and Mean Values of TOTALPD in 5 Datasets

Dataset	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Entire	2.02	169.18	291.50	494.92	648.25	6,849.83
LD	7.39	72.00	83.25	103.28	98.46	1,238.08
PD	2.02	707.12	955.69	1,089.77	1,319.90	6,849.83
OD	29.07	147.51	173.22	215.86	205.15	1,550.00
ND	18.23	231.53	296.24	358.37	424.74	3,153.95

As we mentioned in Subsection 1.1, the maintenance cost is the most significant portion of the basic truck warranty since it covers most systems and parts of the insured truck with a few exceptions, such as routine oil change. The information in this table verifies our original idea to a some extent. Here are two scenarios.

If a payment is in LD, it may imply that the truck does not go through a significant repair. For example, there is a bolt that has become loose (or has gotten lost) and then the garage has tightened this bolt. Since the price of a bolt is low and it is not a complex issue, the total payment will be low, and the proportion of parts payment will be low as well. In addition, the labor payment will dominate the total payment.

If a payment is in PD (the parts payment dominates the total payment), it will imply one or more parts in the truck which be replaced, and since it far exceeds the amount of labor payment, the total payment should be always a large amount.

The features of OD and ND are not as obvious as LD and PD. But the distribution of total payment in OD is quite similar with the distribution in LD. First, the mean value is small relative to PD and ND. The mean value exceeds the third quantile, which means over 75% of payment is below the average, which means small payments dominates these two groups.

2.3.2 Variable 2: deal.to.fail

Next, we are going to verify that there are some relationships between deal.to.fail and four different type of payments. The scatterplots are shown in Figure 1 to Figure 4 in Appendix B.

The average total payment is higher during the second year than the first year, so does the average of each portion of the total payment, especially for parts payment. (See Table 3)

Table 3: Mean Values of Different Payments for the First and the Second Year Records

Year	Labor Payment	Parts Payment	Other Payment	Total Payment
1^{st}	80.02	284.90	109.75	474.65
2^{nd}	88.58	390.28	127.23	606.10

It is worth noting that the average payments are increasing with time goes by, especially for labor payment, parts payment and total payment. So we believe deal.to.fail is a useful indicator to predict the payment amount. And one interesting phenomenon is 9803 payments (84.58% of payments) were incurred within one year after the trucks were sold. The reason for this phenomenon is not clear. But it does not affect our prediction models.

2.3.3 Variable 3: DEFECTCOD and Variable 4: SALESMODL

Finally, we are going to verify that there are some relationships between defect code and total payment. There are two tools to verify: a table to display the number of records in different classes for each DEFECTCOD (see Table A.1 in Appendix A) and boxplots (DEFECTCOD as the variable in x-axis and TOTALPD as the variable in y-axis, see Figure 5 in Appendix B). Take defect code 46 for instance, there are 19 relative records, and 10 of them belong to PD which contains many high payment amounts. The boxplot shows this defect code really has a higher payment amount than many other defect codes. We speculate that this defect code indicates a higher payment amount and we wish to find more relationships between DEFECTCOD and TOTALPD.

The method to verify the existence of relationships between sales model and total payment is similar with the method we mentioned in the previous paragraph (see Table A.2 in Appendix A and Figure 6 in Appendix B). Take EJE225E truck model for instance. From Table A.2, there are 84 relative records, and 42 of them belong to PD and 29 of them belong to ND which contain many high payment amounts. The boxplot shows EJE225E truck model really has a higher average payment amount than many other truck models. We speculate that this truck model indicates a higher payment amount and we wish to find more relationships between SALESMODL and TOTALPD.

CHAPTER 3

MODEL CONSTRUCTION AND COMPARISON

3.1 Framework

In this chapter, we are going to construct 3 models.

Model 1, Model 2 and Model 3 are going to be introduced. Model 1 is our proposed model. Model 2 and Model 3 are constructed for comparison purposes.

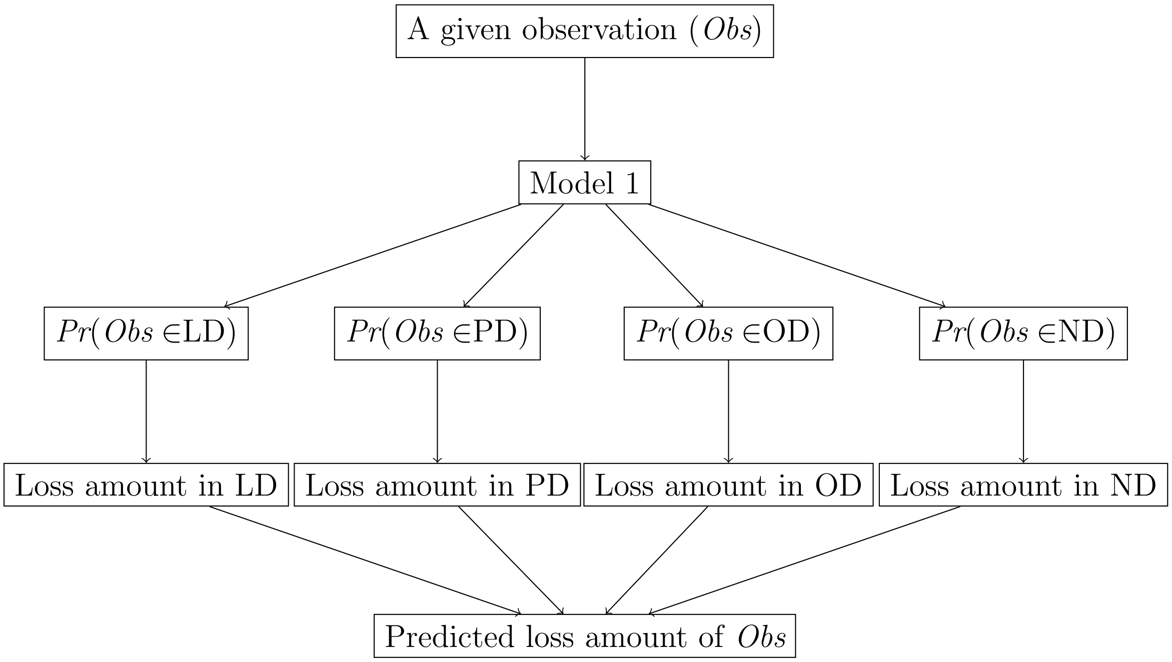
The first model (Model 1) by multiple-class classification is to predict the group number with probability. Then, using total expectation formula, multiplying the probabilities with the mean value of total payments of each corresponding group, we can get a value of total payment as our prediction.

The second model (Model 2, GLM Model) and third model (Model 3, Aggregate Loss Model) are used to compare with Model 1. Model 2 is a regression model and Model 3 is a aggregate loss model. Cross-validation is needed when we construct and validate Model 2. But Model 3 involves payment frequency distribution and severity distribution, and cross-validation destroys the nature of frequency distribution. Therefore, cross-validation will not be applied in Model 3.

3.2 Model 1: Multi-Class Logistic Classification Model

A challenge of our project is that there are only four reasonable variables to construct models and make predictions. Furthermore, two of them contain too many categories, which may cause the overfitting problem when models are constructed if the loss amount is designed to be predicted directly.

In order to better illustrate Model 1, its process is displayed in the following chart.



In the first stage, based on the given data, we can construct a model to predict the probabilities that a given observation belongs to the four groups by the multiple-class logistic regression model. The parameters are fitted by the training dataset. The accuracy of values calculated by confusion matrix will be shown in subsection 3.4.

In the second stage, we can construct a model to predict the loss amount for the given observation, suppose that it belongs to one of these groups. Some advanced algorithms can be applied here, but we use the mean values of payment amount for each group. Not only for simplicity, there are several other reasons. First, the algorithms to predict the loss amount for these four groups should be same. For example, we do not expect the algorithm to predict the loss amount for LD group which is a generalized linear model but the algorithm for OD group is a neural network model. Second, more advanced algorithms may not be able to handle the future situation where the similar situations have not been trained. For example, suppose there is an observation with defect code 4 and truck model ERE225, if the probabilities that the generalized linear regression model are applied to predict the loss amount for the observation, and suppose the predicted probabilities in the first stage are 5%, 10%, 75% and 10%, respectively. However, in our entire dataset, for the LD group, there is no corresponding record for defect code 4 (see Table A.1 in Appendix A) or truck model ERE225 (see Table A.2 in Appendix A). Then, it cannot make a prediction for the loss amount given the observation which comes from the LD group. Third, based on the comparison results (see subsections 3.4 and 3.6), Model 1 has increased the prediction accuracy even if we just use the mean value of total loss amount as the predicted amount.

In the third stage, we find the final value of predicted total payment amount. The third stage can be concisely expressed by total expectation formula, as shown below.

$$\mathbb{E}(\text{TOTALPD}) = \sum_{i=1}^{4} Pr(\text{The record belongs to the Group } i)\mathbb{E}(\text{TOTALPD}|\text{Group } i)$$

On the right side of the formula, we notice that the probabilities are predicted by the logistic multiple-class classification model, where the parameters of the model are fitted by the training dataset. The conditional expectations are estimated by the training dataset based on the average amount. Notice that the training dataset is used for both factors of each term, which means this model takes more advantage of training dataset than other general predictive models such as simple linear regression and other generalized linear models.

3.3 Model 2: Generalized Linear Regression Model

Generalized linear model is a generalization of ordinary linear regression. The error term (which has to be a normal random variable in ordinary linear regression model) could be a non-normal random variable by a link function.[5, 6]

When the response variable is quantitative, a regression model is more likely to be selected than a classification model. Even though the output of logistic regression is quantitative, we always treat it as a classification model since the output is always between 0 and 1.

After we tried several general linear regression models, the normal linear model with the log link on the response variable and the gamma GLM with log link on the response variable can make better predictions than other generalized linear regression models.

Next, we are going to compare the results between Model 1 and Model 2. There are 3 models in Model 2: the normal linear model to predict the logarithm of the response variable (Model 2.1), the normal linear model with the log link on the response variable (Model 2.2), and the gamma GLM with the log link on the response variable (Model 2.3).

3.4 Comparison: Model 1 and Model 2

In order to get rid of overfitting, it is neccessary to divide the dataset into two sets: training dataset and test dataset. In R, the two functions, set.seed and createDataPartition can help us divide the dataset by stratified sampling, but we can divide them identically in different computer as long as the parameter of set.seed remains the same. The training dataset is used to fit the model and get the parameters, and the test dataset is used to validate versatility of the model. We wish the model not only to fit the dataset used to train the model very well, but that it also fits the future dataset decently. This goal could be achieved by showing the prediction on the test dataset is close to the real value of the response variable in the test dataset. In general, the training dataset is 70% to 80% of the entire dataset. In our study, we used 80% of the entire dataset as our training dataset, and the rest as the test dataset.

For some defect codes and some truck models, the relevant records are quite rare, such as defect code 18 (only 3 records) and truck model 2SS2200 (only 4 records). This rarity might cause trouble in the prediction stage. Take the 3 records with respect to defect code 18 for instance. If all of these 3 records are divided into test dataset, then the model constructed by the training dataset does not involve defect code 18. When the model performs the prediction, an error will occur as defect code 18 which cannot be handled by the model. Therefore, we combine all defect codes whose records are fewer than 10 as a new ”code”, and we combine all truck models whose records are fewer than 10 as a new ”truck model” to eliminate the problem we mentioned above.

After the training/test split, we need to verify if the training dataset and test dataset is comparable, because createDataPartition function performs a stratified sampling instead of a random sampling so that both training dataset and test dataset will have representative distributions for the response variable.

The test root mean squared error (test RMSE) is the measurement to evaluate the model, which is defined as follows.

$$\text{test RMSE} = \sqrt{\sum (y_i - \hat{y}_i)^2}$$

where all y_i's come from the test dataset.

Table 4 shows the test root mean squared errors of the 4 models in 20 situations with different training/test splits.

Table 4: Test RMSE of Model 1 and Model 2

	Test RMSE			
set.seed()	Model 1	Model 2.1	Model 2.2	Model 2.3
1	4.2277	142.7698	5.7754	0.8305
2	1.5287	138.0750	0.2281	4.7515
3	2.4457	138.2546	3.9121	0.6742
4	6.6575	135.2902	6.3284	0.9943
5	1.3293	137.8601	0.8280	5.0839
7	0.1604	137.5923	0.7225	4.5268
9	1.7796	138.9027	1.9746	3.0905
10	0.0393	137.9304	3.5081	8.3752
13	2.9950	136.5202	0.0287	0.3786
14	0.1825	135.6620	0.6539	3.6787
15	9.3510	129.1495	7.0690	3.9882
16	3.8706	134.4703	6.7966	0.3487
18	2.9681	135.0606	4.2741	4.4625

19	5.1781	141.9676	4.5476	2.4217
22	2.6676	136.8303	3.2285	5.8111
23	3.6377	138.0106	2.5612	1.3200
24	4.4984	141.2988	3.2275	1.8117
27	0.0647	138.7064	2.8560	1.1203
28	5.1773	138.5588	4.1054	11.222
30	0.9002	138.9466	3.4822	6.3537

Table 5: Information Summary of Table 3

Test RMSE	Model 1	Model 2.2	Model 2.3
Mean	2.9830	3.3054	3.5622
Standard Deviation	2.4362	2.1490	2.8909

The results when the parameter of set.seed function is 6, 8, 11, 12, 17, 20, 21, 25, 26 and 29 were omitted because distributions of TOTALPD in the training set and test set behaves quite differently in these situations and it causes bad performances for all models.

Compared with Model 2.1, the 3 other models improve the accuracy of prediction significantly. Furthermore, from Table 4, we notice Model 1 also improve the prediction accuracy about 10% A more con convincing conclusions can be drawn after hundreds and thousands repeating tests.

The next table shows the prediction accuracy for the fist stage of Model 1 in these 20 situations. The average accuracy is 52.96%. Notice it is a 4-class classification, this accuracy is acceptable, and it shows there is a relationship between different dominated payment types and total payment.

Table 6: Prediction Accuracy

set.seed()	1	2	3	4	5	7	9
Accuracy	0.5281	0.5294	0.5380	0.5354	0.5285	0.5203	0.5320
set.seed()	10	13	14	15	16	18	19
Accuracy	0.5259	0.5332	0.5281	0.5324	0.5246	0.5289	0.5255
set.seed()	22	23	24	27	28	30	
Accuracy	0.5315	0.5376	0.5091	0.5285	0.5281	0.5471	

3.5 Model 3: Aggregate Loss Model

The aggregate loss model is used to evaluate the total loss amount by two aspects of loss: loss frequency and loss severity. Define S, N, and X are the random variables for aggregate loss, frequency, and severity, then

$$\mathbb{E}(S) = \mathbb{E}(N)\ \mathbb{E}(X)$$

and

$$\text{Var}(S) = \mathbb{E}(N)\ \text{Var}(X) + \text{Var}(N)\ \mathbb{E}(X)\ ^2$$

There is no deductible or policy limit in our dataset, therefore, the total payment can be treated as the entire loss for each record.

In general, the frequency distribution and the severity distribution are fitted in order to evaluate their mean values and the variances and preserve for the future use.

Cross-validation cannot be implemented in this model. There might be more than one records for a truck. If cross-validation is implemented, some of these records might be divided into the training set and the rest of records will be divided into the test set. Then the mean value of frequency random variable will be underestimated since only part of real records will be used to train the parameters of this distribution, also, the parameters will be incorrectly estimated. Therefore, we are going to fit the frequency distribution and severity distribution.

As mentioned in section 2.2, there are 15,087 trucks with no payments during the warranty period. Denote it as n_0, which means the number of trucks which has 0 payments during the warranty period. Since we retain the serial number for each record, we could get the number of trucks which has k payments during the warranty period (k is a positive integer). (See Table 7)

Table 7: Empirical Distribution of Payment Frequency

k	0	1	2	3	4	5	6	7	8	9	10
n_k	15087	2839	1138	574	286	168	101	72	53	27	25
k	11	12	13	14	15	16	17	19	22	26	29
n_k	19	6	8	4	6	4	3	1	1	1	1

When fitting the frequency distribution, we first consider distributions from $(a, b, 0)$ class. There is a feature of distributions of $(a, b, 0)$ class [7], which is

$$\frac{n_k}{n_{k-1}} = a + \frac{b}{k} \ (k = 1, 2, 3, ...).$$

Three distributions of $(a, b, 0)$ class are always used to fit the frequency distribution: Poisson distribution, binomial distribution, and negative binomial distribution. However, all these three distributions require a positive b. The scatterplot of $\frac{1}{k}$ v.s. $\frac{n_k}{n_{k-1}}$ implies a negative b. Therefore, $(a, b, 0)$ fails to fit the frequency distribution.

Then, we consider distributions from $(a, b, 1)$ class. Similar with distributions from $(a, b, 0)$ class, distributions from $(a, b, 1)$ class can also make the above equation hold, except for the situation where $k = 2$. That is, for distributions from $(a, b, 1)$ class, we have

$$\frac{n_k}{n_{k-1}} = a + \frac{b}{k} \ (k = 2, 3, 4, ...).$$

Another way to understand distributions from $(a, b, 1)$ class is that they are derived from distributions from $(a, b, 0)$ class. By changing the probabilities at 0 and scaling the other probabilities, we could generate a distribution from $(a, b, 1)$ class.

In our study, we could always focus on the trucks with non-zero number of payments, therefore, we could let the probability at 0 is 0 to make it as a zero-truncated model, and then place a probability at 0 (other probabilities will then be adjusted accordingly), and its maximum likelihood estimator is the frequency at zero in the sample.

Zero-truncated negative binomial distribution could fit the frequency data (with $r = 0.21142$ and $\beta = 2.59855$). Then we need to validate the result by chi-square goodness-of-fit test. The test statistics is 26.20727 with degree of freedom 13. Since the critical value at 5% and 1% significant levels is 24.99579 and 30.57791, respectively, this result can be accepted at 1% significant level.

There are several distributions that are able to fit the severity data (TOTALPD). Notice there are several outliers and these values are much higher than other values, which means a distribution with a heavy tail is better to fit the data than a distribution with light tail. Two typical examples of heavy tail distributions are lognormal distribution and Pareto distribution. However, when we try to use Pareto distribution to fit the data, the result does not converge (different initial values lead to different results). So, we select lognormal distribution to fit the data. By maximum likelihood estimation, the estimated parameters are $\hat{\mu} = 5.759892$ and $\hat{\sigma} = 0.9052144$.

3.6 Comparison: Model 1 and Model 3

The true total payment is 5736083.68. Since there are 20424 trucks in total, the average aggregate payment is 280.85. By total expectation formula, for each truck, the expected payment frequency ($\mathbb{E}(N)$) is $\hat{p_0} \times 0 + (1 - \hat{p_0})\frac{\hat{r}\hat{\beta}}{1-(1+\hat{\beta})^{\hat{r}}} = 0.61$ under the aggregate loss model, where N follows the negative binomial distribution we just fitted and $\hat{p_0} = \frac{15087}{20424}$ is estimated by maximum likelihood estimation. And for each truck, the expected loss severity ($\mathbb{E}(X)$) is $e^{\hat{\mu}+\frac{\hat{\sigma}^2}{2}} = 477.99$, And the expected aggregate for each truck is $\mathbb{E}(S) = \mathbb{E}(N)\ \mathbb{E}(X) = 289.32$. While the average predicted payment by Model 1 is 496.50, given that the truck has at least one payment. On average, the average predicted payment for each truck is $(\frac{11590}{20424}) \times 496.50 = 281.75$. This value is closer to 280.85, so Model 1 outperforms Model 3.

It is worth-noting that the real data is hard to simply be fitted by some classic distributions when we used aggregate loss model to describe the current frequent distribution and severity distribution. For example, the frequency distribution cannot be accepted at the 5% significant level. In this respect, Model 3 does not fit this dataset very well.

CHAPTER 4

CONCLUSION AND FUTURE WORK

4.1 Conclusion

Since there is no price discrimination on basic truck warranties for new trucks, we expect a more accurate prediction for the average total payment amount. The comparisons with generalized linear regression models and aggregate loss models show the model we constructed can achieve that goal. Moreover, compared with generalized linear regression models, we notice our model can make a more accurate prediction not only for the average total payment amount, but also for the total payment amount for each record, which means this model far exceeds our original expectations. However, this is a data-driven model, which means the result highly depends on the data. The versatility could be verified by more similar types of data. And there could be more improvements which will be discussed in the next subsection.

4.2 Future Work

This model is relatively straightforward. New techniques and ideas can make it more advanced. Due to the time restriction, we will only discuss several possibilities of improvements rather than implement them.

First, we did not use all variables we selected. For example, we might take more advantage of DEALERCOD (the numerical ID of each dealer). Then a more complex model could tell us which dealer could sell trucks that have good quality and workmanship versus those that are not. As a research topic for future work, more advanced techniques such as BERT algorithm [8] in Nature Language Processing (NLP) could be applied to predict the payment amount. For example, by NLP techniques, the cause of each payment could be extracted and transformed from descriptive information to quantitative information. Then it could be treated an important factor to improve the accuracy of the prediction, similar to the review helpfulness studies [9]. This model is a data-driven model, which means the model is quite subject to the dataset. Changing another set of data may make this model invalid.

Second, this model is not good at predicting large amount of payments. Since the predicted probabilities are used in the total expectation formula with the mean values of total payment in four classes. The largest predicted value is 1089.77 (when the predicted probability in PD is 1). Actually, there are 1333 (11.50%) records having total payment larger than 1089.77.

Third, it might not be fair to charge the same warranty premium for different models. In this dataset, we did not take that into our consideration because we do not have enough records to train a model for each truck model. To solve this problem, more records could be collected, or bootstrapping method could be applied to generate more simulated records.

BIBLIOGRAPHY

[1] Brown, Robert L., and Leon R. Gottlieb. *Introduction to ratemaking and loss reserving for property and casualty insurance.* Actex Publications, 2007. 434-447.

[2] Fu, Luyang, and Hongyuan Wang. *"Estimating insurance attrition using survival analysis."* Variance 8:1, 2014, pp. 55-72.

[3] Duncan, Ian, et al. *"Using Survival Analysis to Predict Workers' Compensation Termination."* Variance 13:1, 2020, pp. 31-53.

[4] James, Gareth, et al. *An introduction to statistical learning.* Vol. 112. New York: springer, 2013. 135-137.

[5] Gross, Chris, and Jon Evans. *"Minimum Bias, Generalized Linear Models, and Credibility in the Context of Predictive Modeling."* Variance 12:1, 2018, pp. 13-38.

[6] Spedicato, Giorgio Alfredo, Christophe Dutang, and Leonardo Petrini. *"Machine learning methods to perform pricing optimization. A comparison with standard GLMs."* Variance 12.1, 2018, pp. 69-89.

[7] Klugman, Stuart A., Harry H. Panjer, and Gordon E. Willmot. *Loss models: from data to decisions.* Vol. 715. John Wiley Sons, 2012. 83-88.

[8] Devlin, Jacob, et al. *Bert: Pre-training of deep bidirectional transformers for language understanding.*, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies," Volume 1, 2019, pp. 4171–4186.

[9] Xu, Shuzhe, Salvador E. Barbosa, and Don Hong. *"Bert feature based model for predicting the helpfulness scores of online customers reviews."* Future of Information and Communication Conference. Springer, Cham, 2020, pp.270-281.

APPENDICES

APPENDIX A

LONG TABLES

The following table shows the class distribution for each defect code.

Table A.1: Class Distribution for DEFECTCOD

	Class					Class			
DEFECTCOD	LD	OD	ND	PD	DEFECTCOD	LD	OD	ND	PD
1	0	0	8	1	51	13	36	216	95
2	1	0	5	3	52	5	2	100	28
3	0	1	8	0	54	7	6	11	4
4	0	2	3	2	55	0	2	1	3
5	0	2	12	6	57	8	2	9	0
6	3	6	38	52	58	44	141	134	12
10	123	282	672	128	59	2	3	22	5
11	17	35	65	77	60	17	3	10	5
12	34	4	36	7	62	1	1	4	2
15	1	3	10	3	63	1	1	0	0
16	0	0	5	1	64	1	2	3	1
17	49	93	275	80	66	0	1	3	2
18	0	2	1	0	69	26	3	13	16
19	0	1	0	3	70	0	12	17	10
21	3	1	23	8	71	1	0	1	0
22	49	102	122	20	72	0	10	14	8
23	9	19	166	22	73	1	9	36	27
24	0	16	65	0	76	0	2	2	2
25	2	1	10	2	77	3	0	1	1
27	4	7	35	2	78	30	88	354	398
31	29	257	434	644	79	47	27	13	7
32	47	414	461	264	81	0	19	31	41
33	5	14	22	23	84	0	0	2	1
34	2	15	29	8	85	4	8	23	0
35	0	0	1	1	86	0	1	0	0
36	7	17	5	54	87	1	5	19	0
37	0	4	2	2	88	0	0	4	1
39	0	0	1	1	89	586	3	75	2
40	1	10	4	32	90	1	1	10	0
41	2	11	3	12	91	127	46	34	6
42	124	263	481	857	92	0	0	8	1
43	1	4	14	6	93	127	231	208	85
44	0	0	3	2	96	0	1	1	5
45	8	10	36	32	97	3	6	22	11
46	2	5	2	10	98	0	0	0	1
47	5	17	21	4	101	2	0	0	0
48	0	4	8	8	N/A	4	15	37	11

The following table shows the class distribution for each truck model.

Table A.2: Class Distribution for SALESMODL

	Class					Class			
SALESMODL	LD	OD	ND	PD	SALESMODL	LD	OD	ND	PD
2SS2200	2	1	0	1	PMWF11N	0	1	2	3
2WP4500	331	439	408	364	PMWR30N	13	52	106	91
2WP6000	6	24	26	23	PMWR30NQ	0	1	7	0
2WPL4500	399	420	634	431	PMWR40N	0	1	3	1
2WPL6000	44	60	118	51	PMWT11N	0	0	1	2
ECR327	18	67	121	59	PMWT15N	0	2	2	0
ECR336	16	22	31	3	PMWT18N	0	0	4	1
ECR360	2	1	0	0	PW23	5	113	173	135
EJE120	94	388	628	650	PW23L	0	0	0	4
EJE120E	9	12	32	16	PW30	0	2	4	3
EJE225	2	4	8	14	PW30L	1	24	94	57
EJE225E	5	8	29	42	SF2200	2	9	12	19
ERE225	0	2	6	3	SS3000	7	3	9	3
ESE120	0	0	0	1	SS3500	1	3	10	3
NSP12N2	0	0	1	0	WP3000	0	0	1	2
NSP16N	0	0	1	5	WP4500	48	97	146	99
NSP16N2	0	0	1	0	WP4500L	65	107	104	116
PMW15N	0	2	1	1	WP6000	2	5	13	6
PMW23N	2	13	34	13	WP6000L	4	35	80	78
PMW23NL	1	1	3	1	WR6000	507	378	1642	854
PMW23NQ	0	2	5	4	WR8000	4	8	23	9
PMW30N	0	1	1	0	N/A	0	0	0	0

APPENDIX B

FIGURES

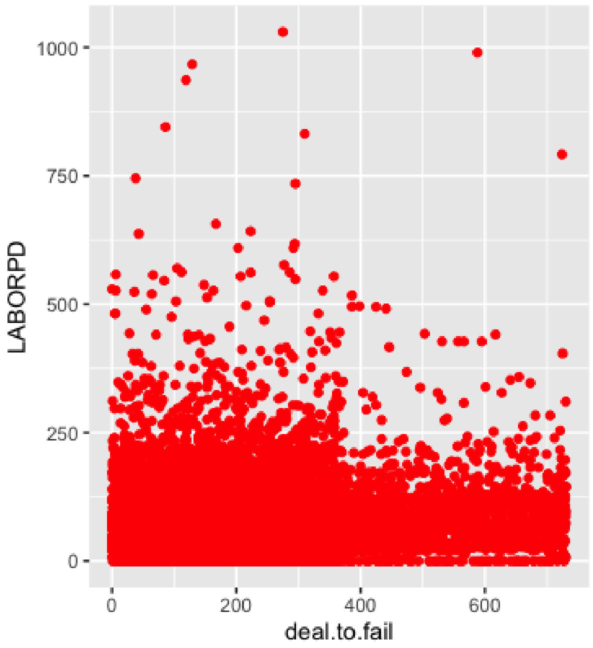


Figure 1: Scatterplot between deal.to.fail and LABORPD

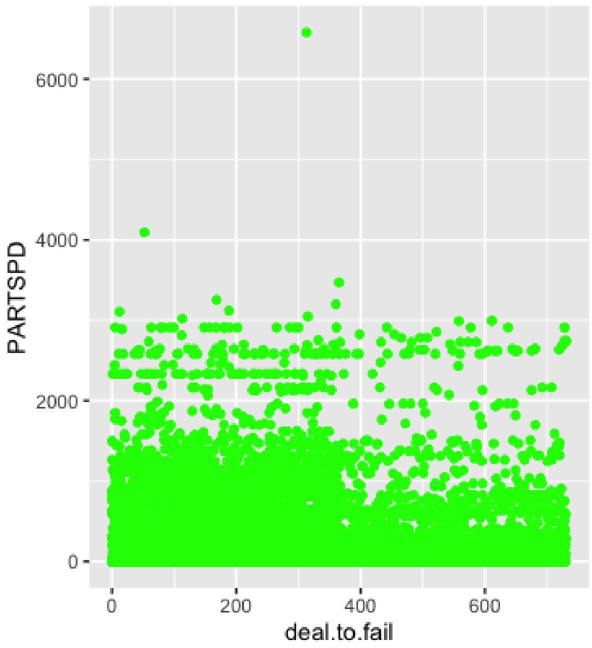


Figure 2: Scatterplot between deal.to.fail and PARTSPD

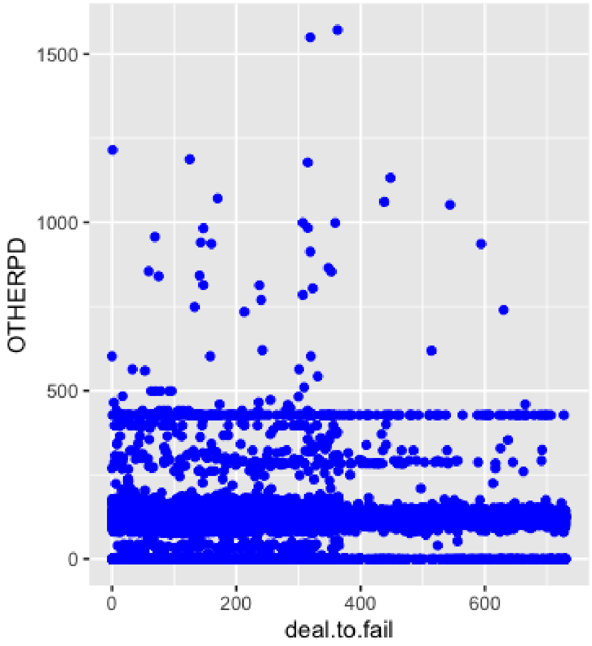


Figure 3: Scatterplot between deal.to.fail and OTHERPD

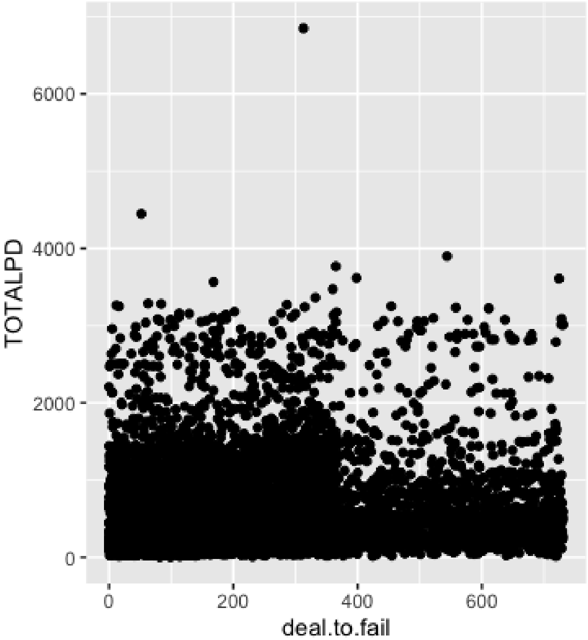


Figure 4: Scatterplot between deal.to.fail and TOTALPD

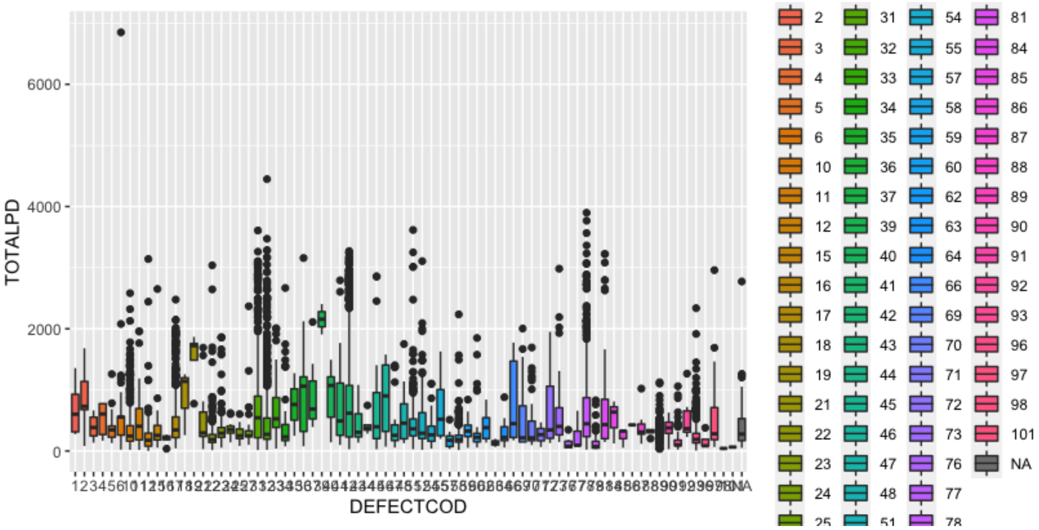


Figure 5: Boxplot between DEFECTCOD and TOTALPD

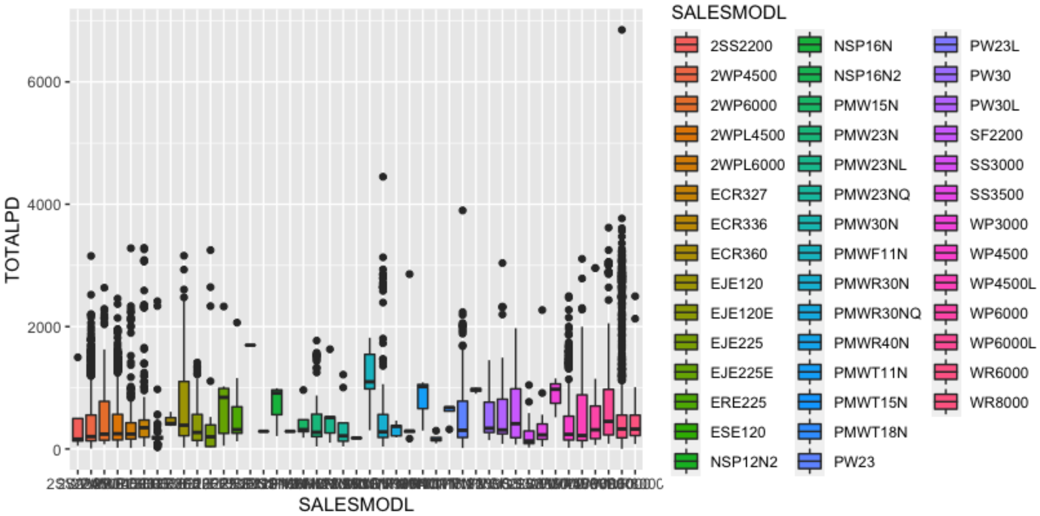


Figure 6: Boxplot between SALESMODL and TOTALPD