Data Science Team Project: Customer Lifetime Value Analysis

Ziyi (Summer) Fu, Chunyu (Sarah) Ji, Aastha Thakkar, Ning Xu, Yifan Zhou

Introduction

Our core task for this report is to use supervised prediction models and our target is the Customer Lifetime Value($) for an Auto Insurance company. We use data mining methods such as linear regression, stepwise regression, neural nets, support vector machine, lasso and post lasso to support this analysis. We use K-fold Cross Validation to pick the best final model, and evaluate its robustness based on the out of sample R^2 from the testing dataset.

Business Understanding

Our report analyzes data from an auto insurance company. Insurance is a complicated industry and a company has a fair amount of discretion on how it chooses to process claims. Given the lack of transparency in the insurance industry, a company can actually use Customer Lifetime Value (CLV) to guide its claim payouts. Simply trying to focus on a high CLV might lead to accepting too many risky customers, but using CLV as a guide for customer service and payouts will ultimately lead to higher profitability.

We decided to use CLV as our main target is that CLV is a useful matrix for cost-benefit analysis. The company can compare the projected CLV to the costs of acquisition and retention to determine long run profitability. They can design offers based on each customer's CLV. Do a CLV analysis by product line and determine a better mix of offerings or by sales channel for internal evaluation. CLV can also help the company guide customer service. The company can differentiate customers by making the application more complicated for less valuable customers while still keeping profitable customers engaged. Those differentiated groups can also be used to inform future acquisition efforts.

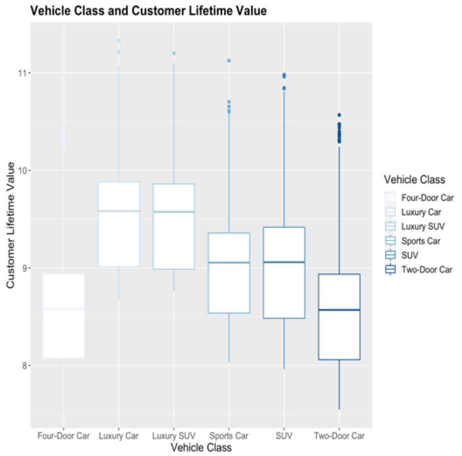
To increase CLV, an auto insurance company can intuitively maximize its premium payments. The company can also target multiple policies to gain higher CLV. Sales channels might also affect CLV in a way that customers might compare different brands and ‘bargain’ through online channels so the acquisition costs will be higher. In terms of customer profiles, the employed and educated customers might have higher CLV compared to unemployed and less educated ones. The class difference in their vehicles might affect CLV a lot as well.

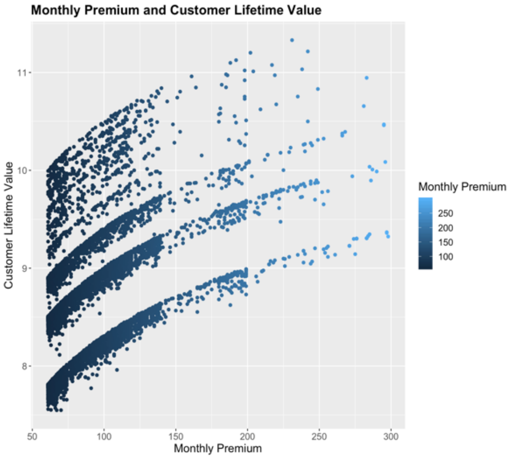
Our regression solution can address the prediction of CLV by determining the importance of the different components of a customer profile that lead to a high CLV. We can retain and acquire customers that already meet those criteria and churn the customers who do not meet them.

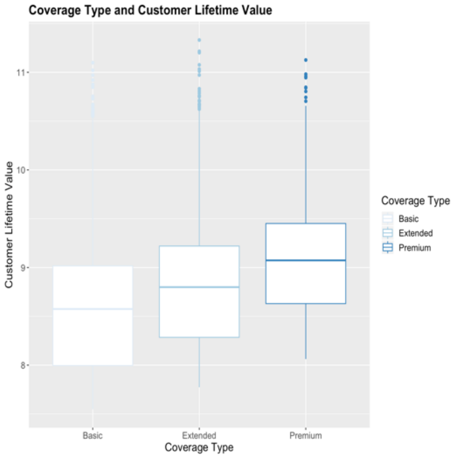
Data Understanding And Exploration

Our dataset was retrieved from *kaggle*. It has 24 variables describing insurance policies and customer information in an auto insurance company. There are 9134 records in this dataset. Our dependent variable is CLV, and most of the other variables are unordered categorical variables such as policy type, gender, education, etc. which reveal customers’ characteristics. Our analysis aims to explore the drivers behind CLV.

Exploratory Data Analysis



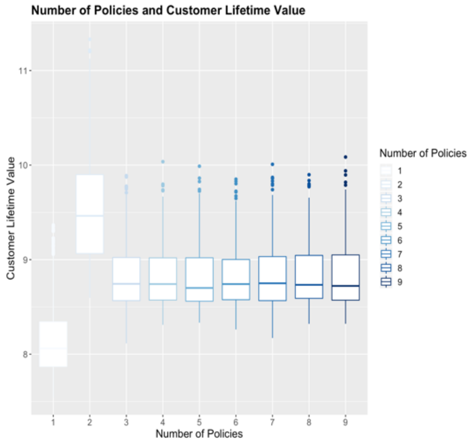


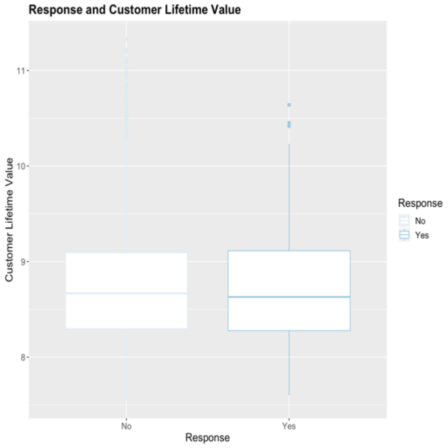


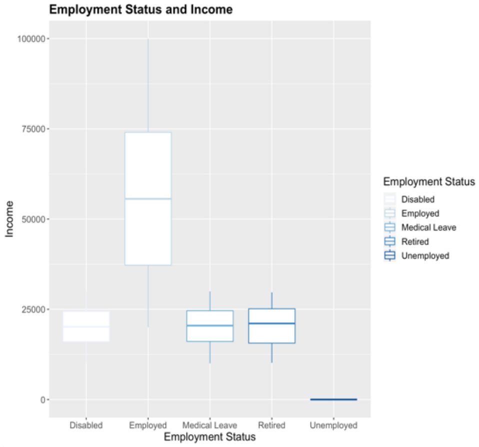
I. **Vehicle Class vs CLV:** Different types of vehicle owners have different value to the company. We noticed that luxury auto owners tend to have higher CLV, followed by SUV and sports car owners.

II. **Monthly Premium vs CLV:** CLV tends to be highest when customers pay a higher monthly premium.

III. **Coverage Type vs CLV:** Coverage type can also affect customer lifetime value. Higher level of coverage typically brings more value.



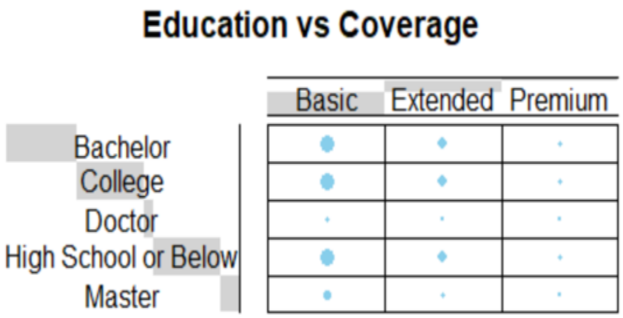




IV. **Number of Policies vs CLV:** We notice that 2 policies is ideal to maximize customer lifetime value. Introducing only one policy is not efficient, but forcing more number of policies might cause customers to leave.

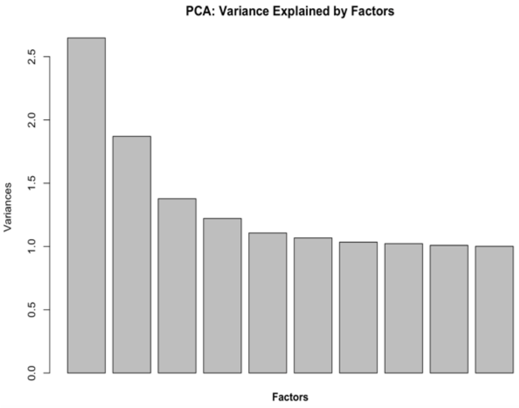
V. **Response vs CLV:** While we do not know how customers respond to future renewal offers, it is possible that CLV is higher if a customer responds yes to a renewal offer. However, we did a biserial correlation test, and the point-biserial correlation is smaller than 0.2, meaning they are independent; so we will not take Response into account.

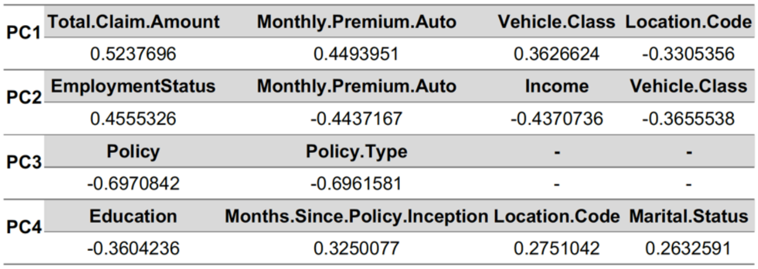
VI. **Employment Status vs Income:** We noticed that a quarter of observations in our data have 0 income. This matches exactly with everyone who is unemployed.



VII. **Education vs Coverage:** We also plot contingency tables for categorical variables like Education and Coverage. The grey area represents the correlation. We find relatively high correlation exists in between these two.

Principal Component Analysis





PC1	Total.Claim.Amount	Monthly.Premium.Auto	Vehicle.Class	Location.Code
	0.5237696	0.4493951	0.3626624	-0.3305356
PC2	EmploymentStatus	Monthly.Premium.Auto	Income	Vehicle.Class
	0.4555326	-0.4437167	-0.4370736	-0.3655538
PC3	Policy	Policy.Type	-	-
	-0.6970842	-0.6961581	-	-
PC4	Education	Months.Since.Policy.Inception	Location.Code	Marital.Status
	-0.3604236	0.3250077	0.2751042	0.2632591

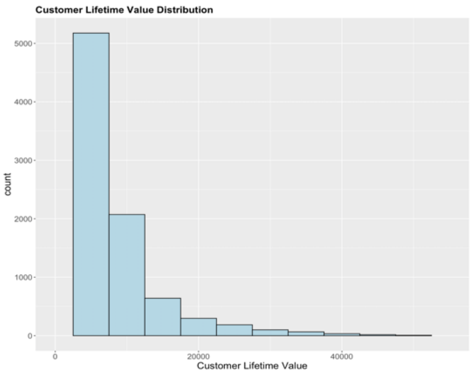
After transferring all the categorical variables to numeric variables, we are able to conduct a Principal Component Analysis to help us explore potential patterns in the dataset. We then pick the first 4 components to interpret the composition of each group.

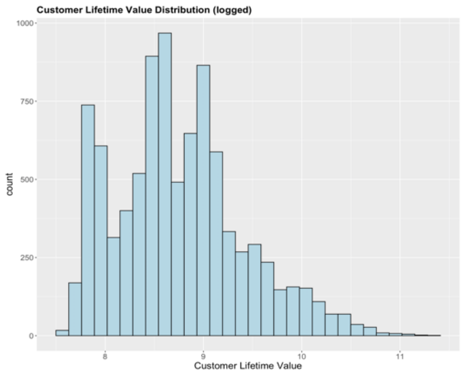
Looking at the first PC, we can tell this group has customers who are more likely to have high cost accidents, higher premiums, and high end cars. The second PC has customers who pay lower premiums, and are usually employed with lower incomes. The third PC has customers signed up for corporate insurance policies and the last PC are customers who are less educated, single and own lower end vehicles.

This method is effective for dimension reduction and we could do further analysis based on 4 significant PCs. However, since most of the data is unordered categorical variables and some information like Total Claim Amount are unavailable when customers first sign up, we are unable to move onwards with the current dataset. Also, the correlations among different PC profiles and variable will be numerous and cumbersome to interpret. Most importantly, given we have enough sample data with CLV and that supervised learning is considered superior to unsupervised in terms of predictive modelling, in this analysis we decided to utilize different models in supervised learning.

Data Preparation

First, we make a histogram to see the overall distribution of the dependent variable, and we notice the deviation pattern. So we logged customer lifetime value and get distribution closer to normal distribution.





Variables we have when a new customer signs up include state, coverage, education, employment status, gender, income, location code, marital status, monthly premium, number of policies, policy type, policy, sales channel, vehicle class, vehicle size. We dropped other variables that we do not have at the beginning such as total claim amount, offer type, response to offer.

We divided the data set into training set and test set based on the 80-20 rule of thumb. By using similar data for training and testing, we can minimize the effects of data discrepancies and better understand the characteristics of the model. After a model has been processed by using the training set, we can evaluate the model by making predictions against the test set. We begin analysis on the training set based on variables from our exploration.

Modeling

In this part, we will build several different models such as a basic model, interaction model, stepwise selection model, Neural Network, SVM, Lasso and Post Lasso to predict CLV. We will compare these models and select the best one in terms of out of sample R^2 and root mean square error (RMSE).

Null Model And Interaction Model

After data exploration, we keep 16 variables as independent variables and CLV as our dependent variable. The null model has an R^2 of 24.54%. In the following discussion, we will use this model as a baseline for later comparison. This forms our null model shown below:

Null Model: *log(Customer.Lifetime.Value) = State + Coverage + Education + EmploymentStatus + Gender + Income + Location.Code + Marital.Status + Monthly.Premium.Auto + Policy.Type + Policy + Months.Since.Policy.Inception + Number.of.Policies + Vehicle.Class + Vehicle.Size + Sales.Channel*

In the data exploration part, we noticed that coverage and education have a high correlation, so we included this interaction in the model and we notice that the adjusted R^2 slightly increased from 24.54% to 24.58%, and the interaction is significant.

Interaction Model: *log(Customer.Lifetime.Value) = Coverage * Education + State + Coverage + Education + EmploymentStatus + Gender + Income + Location.Code + Marital.Status + Monthly.Premium.Auto + Policy.Type + Policy + Months.Since.Policy.Inception + Number.of.Policies + Vehicle.Class + Vehicle.Size + Sales.Channel*

Stepwise Variables Selection Model

We then use stepwise variable selection to predict the CLV. The model is listed below, with an R^2 of 24.53%.

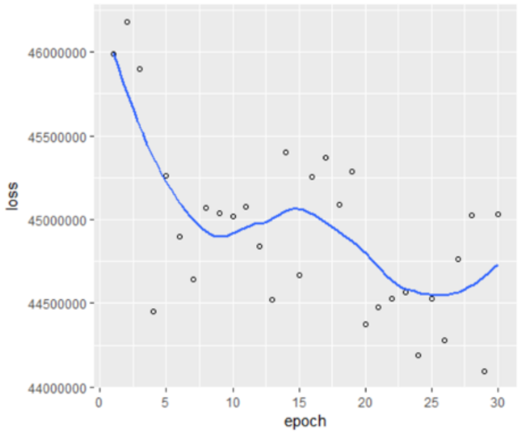
Stepwise Model: *log(Customer.Lifetime.Value) = Coverage + EmploymentStatus + Marital.Status + Monthly.Premium.Auto + Number.of.Policies + Vehicle.Class + Vehicle.Size + Policy.Type*

Neural Network

Then we apply a neural network model to predict CLV. We use keras to build the neural model and put the training data into it to get the predicted values for CLV. Furthermore, in the process of building the neural network, we adjust the parameters repeatedly and yield a relatively good model. Then since we do not have enough variables and observations to prevent overfitting and gradient vanishing, we choose two layers.

To evaluate the results, we compare the neural network with the null regression model above. We calculate rmse for actual CLV with predicted values from neural network and the regression model respectively. We evaluated the neural model in training and testing set. Below is the loss plot for the training set. We can observe that as the number of epoch increases, the loss tends to decrease. Besides loss, we find rmse for actual data and predicted neural is 6443, which is smaller than rmse of 6561 for actual data and predicted regression. This means our neural network model is slightly better than the null model in the training set.

Hence, we believe the predicted CLV from neural network performs as accurately as the null model, and we decide to move forward and compare the results with other models.



Support Vector Machine

We use SVM model to predict CLV as our next choice. After several trials, we pick Employment Status, Coverage Type, and Monthly Premium in the model. From the output, we notice that employed customers can bring higher value to the company and customers who pay higher monthly premium are more valuable as well. In terms of coverage type, we notice a slightly higher CLV for those with premium coverage. The R^2 for this model is 12.6%, so SVM performs better than the dummy average model.

Lasso And Post Lasso

We then run lasso and post lasso to get the models. Lasso can help improve the prediction accuracy and interpretability of regression models by fitting and penalizing coefficients of variables. For lasso regression, sports car and SUV are still shown as important variables. When all else held equal, owning a sports car and owning SUV tend to increase CLV by 13% and 14% respectively.

Lasso min: *log(Customer.Lifetime.Value) = State + Coverage + Education + EmploymentStatus + Location.Code + Marital.Status + Monthly.Premium.Auto + Number.of.Policies + Policy + Sales.Channel + Vehicle.Class + Policy.Type + Coverage* Education*

Lasso 1se: *log(Customer.Lifetime.Value) = EmploymentStatus + Marital.Status + Monthly.Premium.Auto + Number.of.Policies + Vehicle.Class*

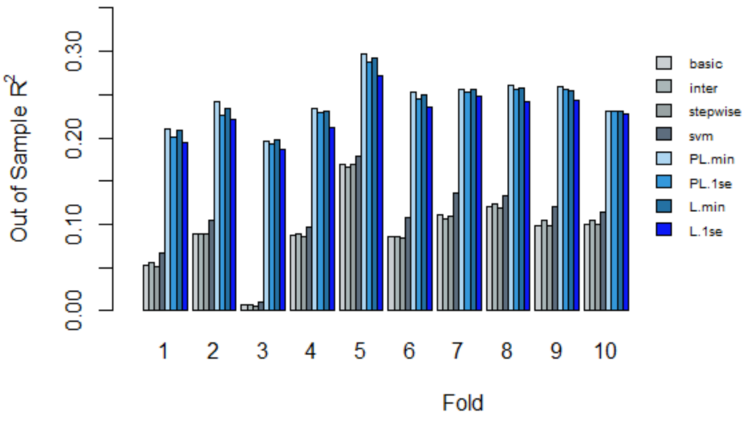
Post lasso min: *log(Customer.Lifetime.Value) = Coverage + Education + EmploymentStatus + Gender + Location.Code + Marital.Status + Monthly.Premium.Auto + Number.of.Policies + Policy + Sales.Channel + Vehicle.Class + Policy.Type + Coverage* Education*

Post lasso 1se: *log(Customer.Lifetime.Value) = EmploymentStatus + Monthly.Premium.Auto + Number.of.Policies + Vehicle.Class*

Compared with lasso regression, post lasso drops the State variable, and includes the Gender variable in the model. The coefficients for sports car and SUV increase. Furthermore, people with Premium coverage and Master education background are likely to have a higher CLV by 0.15% than those with Extended coverage and College education background.

Evaluation

In total, we now have 9 models that can be used to predict CLV. In order to choose the best model, we run K-fold cross validation to test their out of sample (OOS) R^2, which is a good measurement of whether the predictive relationship has out-of-sample predictability. We evaluate the models in terms of bias and variance.



From the k-fold validation graph on the left, we can easily observe two groups of models. Basic model, interaction model, stepwise model and svm model share similar patterns with each other. They have low OOS R^2 (high bias) and high variance, which signals poor performance. In fold 3, their OOS R^{2}s are nearly 0.

In comparison, the patterns for lassos and post lassos are very different. The variances for each model in this group are relatively small. In terms of OOS R^2, their average scores are around 0.25, much higher than the ones from the other group (OOS R^2 ≈ 0.10). Although 0.25 is still not a very high value in terms of OOS R^2 (high bias), there is normally trade-offs between low variance and high bias. Therefore, in general, this group of models demonstrate better performance. One thing we should notice is that most models we have on the graph are significantly better than the basic regression model in terms of out of sample R^2. In retrospective, our neural network model is almost as good as the basic regression model in terms of rmse, which means our predicted CLV values from the neural network should not be used for its lack of accuracy. Therefore, after comparison of 9 models, we decide our final model to be post-lasso min, which has the highest average OOS R^2. Below is the table of variable coefficients for our final model.

Variables	Coef.	Variables	Coef.
(Intercept)	7.880197686	**CoverageExtended**	0.072580753
CoveragePremium	0.033315462	**EducationDoctor**	-0.091396372
EducationHigh.School.or.Below	0.021496465	**EmploymentStatusEmployed**	0.039811799
EmploymentStatusUnemployed	-0.025918476	**GenderM**	-0.010240720
Location.CodeSuburban	-0.022849998	**Marital.StatusSingle**	-0.056867617
Monthly.Premium.Auto	0.006951395	**Number.of.Policies**	0.055290733
Policy.TypeSpecial.Auto	0.071051861	**PolicyCorporate.L2**	-0.041999863
Sales.ChannelWeb	-0.033311125	**Vehicle.ClassSports.Car**	0.174014377
Vehicle.ClassSUV	0.170092574	**CoveragePremium.EduCollege**	0.099385267
CoveragePremium.EduDoctor	0.206585813	**CoveragePremium.EduHigh.School.or.Below**	0.089216540
CoveragePremium.EduMaster	0.150562662		

From the table, we can tell that vehicle class, coverage, coverage*education, marital status, number of policies and policy type special auto are all important variables to post-lasso min model. When all else held equal, people with sports car and those who have SUV on average tend to increase their CLV by 0.17%. People with Premium coverage and Doctor education background are likely to have a higher CLV by 0.20% than those with Extended coverage and College education background. People who are single are likely to decrease their CLV by -0.06%.

We then apply post-lasso min model to the test dataset and run the prediction. We compare our predictions with true values, and find the R^2 equals 27.06%, which matches the OOS R^2 in cross validation in train dataset.

Deployment

Our final CLV prediction model allows us to observe significant drivers for higher CLV: a higher monthly premium, special policies, higher education, sports or luxury cars. Therefore, the insurance company can gain more revenue incentives by targeting educated, luxury car owners for premium policies.

However, the temptation while deploying this CLV model would be to ignore the scale. This is a fairly narrow pool of customers. 9134 observations might suffice for the other models but might not be ideal for neural nets. The norm between ‘quantity’ and ‘quality’ is important. As other companies can also be targeting those premium auto owners, our company can gain competitive advantage via extended coverage users, since the latter also leads to high CLVs and 30% of customers use extended while only 9% use premium.

The key issue for deployment will be to make sure the business does not blindly focus on maximizing CLV; it should instead use this to guide its customer service and payout policies to ensure that the company is not spending more on retaining customers than the value they are gaining from them. For example, while a web sales channel leads to customers with lower CLVs, but it may also be cheaper to run than hiring a field sales team. Corporate policies also generate a lower CLV, but those bulk policies are easier to process than individual ones.

Profitability is also followed by ethical issues. Customers with low CLVs are usually customers with lower incomes or those who are unemployed, so it could be considered unethical to process their claims more stringently than others. Another ethical consideration could be data privacy. It is important to have the right cybersecurity measures to secure the data. Even if selling the data could be a revenue source, it is important to know that customers have a reasonable expectation to privacy and their information must be secure.

A potential risk is that this model is based on information that is almost entirely self-disclosed. People may be tempted to lie about their employment status or income if they believe it will get them more favorable premiums. Another risk is that the training data was already cleaned and this may have led to loss of information. A removed outlier could be dangerous for an insurance company as it means they miss out on calculating the worst case scenario and owe a payout much higher than the CLV. This risk can be mitigated by running this analysis with more of the original data, and the analysis in general can be improved by collecting more granular demographic data like zipcode or age.

Appendix

Ning Xu: Ning's contribution consists of data preparation and the neural network model. To be more specific, data preparation includes variable selection, table joining, creating training and test sets, and transforming unordered categorical variables into dummy binary variables. What's more, works in the neural network model includes building the structure of the model, adjusting the parameters, and evaluating predicted CLV in training and test set respectively.

Summer Fu: Helped draft and polish 'Introduction', 'Business Understanding' and 'Deployment' part in the final report. Led extensive discussion on the project in terms of the business content and structure. For the data processing purpose, Summer constructed lasso and post lasso model for CLV prediction, which are used in the K-fold Cross Validation and OOS analysis later. Lastly, when most of the content is written down, Summer did the formatting for the whole report to make it more consistent in each part and more readable.

Aastha Thakkar: Found the dataset and helped decide on the target variable. Researched the auto insurance industry understand variable selection, to write the business intuition and deployment portion. Collaborated on data explorations and visualizations. Worked on overall coordination ensuring consistency of the report in compiling it.

Chunyu (Sarah) Ji: Sarah explored the relationship among different categorical variables and determined which interactions to add to the model. She was mainly responsible for the modelling building and evaluation part, as well as the writing for modelling. She built and interpreted 7 models, including basic model, interaction model, stepwise selection model, 2 lasso and 2 post lasso models. She also ran K-fold cross validation for all the models and chose the best model based on out of sample R2. After the final model selection, she applied the best model to test dataset for prediction and evaluate its performance.

Yifan Zhou: Yifan is responsible for data visualization, SVM modeling, and R-file compilation. Firstly, she looked into the dataset and connected different variables in variable selection. Moreover, she explored several data mining methods including neural network model and K-means clustering. Then she constructed SVM model and collaborated in different model constructions. Finally, Yifan compiled the R file and did debugging work.

Generally speaking, everyone had their own strengths to bring to the table and overall it was a collaborative team effort with every team member verifying every section to make sure we productively discussed and agreed on everything we submitted.