Section A, Team 10

Irene Guo, Lucy Meng, Hasan Muhammad, Manavv Shah, Tony Qiu

Team Project Final Report: Predicting Interest in Vehicle Insurance

Business Understanding

Our business problem involves supporting a health insurance company that is interested in offering vehicle insurance to its existing customers. We plan to help the company identify which customers to target for their vehicle insurance marketing campaign. Specifically, we will analyze features of the company's existing healthcare customers to estimate the probability that an existing customer is interested in vehicle insurance. Knowing which customers to target is important so that the company does not waste money and resources by advertising to customers that are highly unlikely to be interested. We plan to use predictive modeling to calculate the probability that an existing healthcare customer is interested in vehicle insurance. We plan to test supervised learning techniques such as logistic regression and classification tree to compute the probability of interest of a customer and to validate these models through cross-validation to identify the best model. With these probabilities, the company can perform a cost-benefit analysis for each customer to assess if the value of advertising to a specific customer exceeds the cost of advertising. The probability threshold at which the company decides to market to a customer will vary from customer to customer as it will depend on the anticipated additional premium being charged to that customer if the customer is interested and subsequently proceeds to obtain vehicle insurance. The company would be able to assess which customers to target and market to and would then be able to plan out its communication strategy to reach out to relevant customers, to maximize its profit through an optimal advertising strategy.

Data Understanding

The dataset provided is at the customer level and contains information on various customer features such as gender, age, and vintage (how long the customer has been insured by this company). The dataset also includes features that are useful when analyzing vehicle insurance, such as vehicle age, whether the vehicle has been damaged in the past and whether the customer already has existing vehicle insurance. The target variable (dependent variable) is Response, which is a binary variable that has the value of 1 if the customer is interested in vehicle insurance and 0 if the customer is not interested. The dataset does seem to be imbalanced, as only 12.26% of existing customers are interested in obtaining vehicle insurance from the company. The data also has an element of bias as it consists of the company's existing customers and not new customers. To gain an understanding of which variables might impact the interest of a customer, we generated a correlation matrix plot.

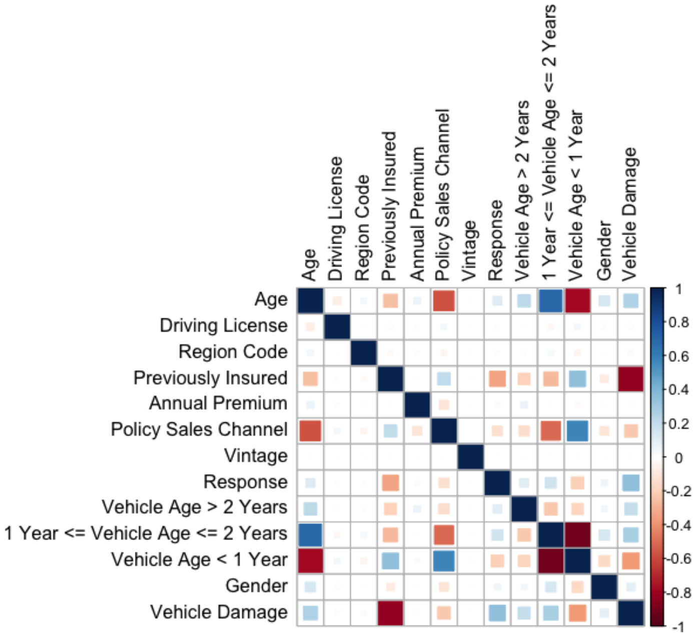


Figure 1: Correlation Matrix Plot

We notice a negative correlation between previously insured variables and response, indicating that customers with vehicle insurance are unlikely to be interested in switching their vehicle insurance to this company. We noticed a positive correlation between vehicle damage and response, which suggests that customers who have had vehicle damage in the past are likely to be interested. There seems to be almost no correlation between vintage and response which suggests that customer loyalty to health insurance does not necessarily translate into customer interest in vehicle insurance. We then proceed with a deep dive into the two most correlated variables with response: Vehicle Damage and Previously insured. The bar plot below shows that the most interest is demonstrated by customers that currently do not have any vehicle insurance and have had vehicle damage in the past. This reflects a key segment that the company can focus its targeting efforts on.

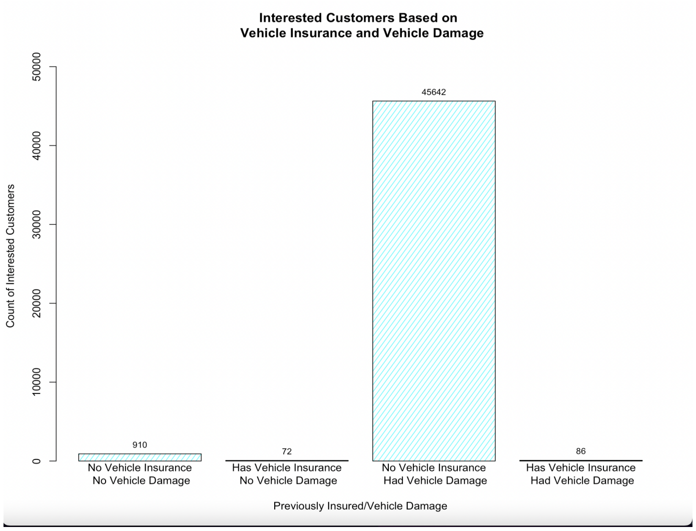
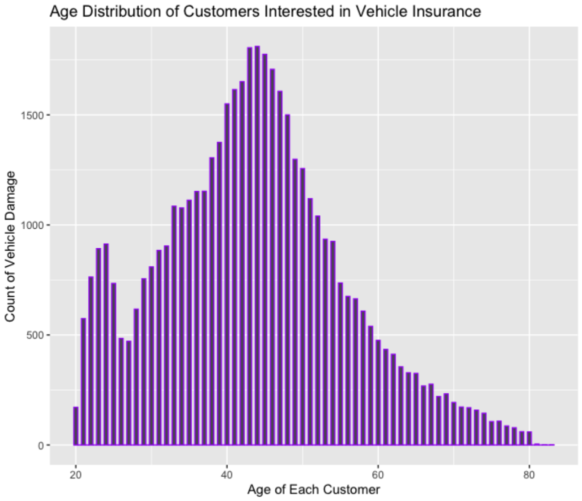


Figure 2: Categories of people more likely to take vehicle insurance

We then proceeded to visualize age distribution with the count of vehicle damage, as both these features could be relevant for the company to price their additional premiums. We notice that customers in their early 20’s and 40‘s had a higher chance of vehicle damage in the past. We also noted that there does not seem to be any strong relationship between the age of the customers and their existing annual premiums.



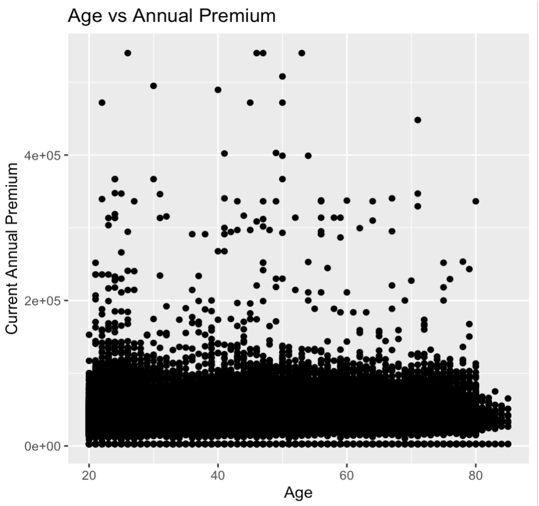


Figure 3: Age Distribution of Customers with Vehicle Damage

Figure 4: Age vs Annual Premium of Customers

Data Preparation

Our dataset contained multiple categorical variables, of which we had to transform the following three: Gender, Vehicle_Damage and Vehicle Age. The latter of the 3 listed an age range rather than the actual age. Gender and Vehicle Damage were binary variables for which we created dummy variables. For Vehicle Age, we created dummy variables for the three categories (< 1 year, 1 - 2 years, > 2 years). The remaining categorical variables, which consisted of previously insured, driver-licensed, and response already had binary values of 0 and 1 so no further transformation was needed. We ran a summary of the dataset to check for any missing values. No missing values were noted and so we did not need to drop any rows or impute any values. We also reviewed the minimum and maximum values of each column to ensure that values were in an appropriate range. We noted that all the existing customers were in the age range of 20 - 85, and the vintage (number of days the customer has been with the firm) of each customer was in the range of 10-299.

Modeling

Our business problem involves modeling the probability that a customer would be interested in vehicle insurance. As our target variable (response) is binary, which can take a value of 1 to indicate the customer is interested in vehicle insurance and take a value of 0 to indicate the customer is not interested, we proceeded with a predictive modeling approach, creating models to predict the likelihood that a customer is interested. We created and tested nine different models: Logistic Regression, Logistic Regression with Interaction Terms, Classification Tree, three models with Lasso - each with a different value of lambda, and three models with Post Lasso based on the variables identified in the Lasso models. The Lasso and Post Lasso models used a subset of available variables as these models were intended to assess the most relevant variables for prediction, while the other models leveraged all variables provided in the dataset. We compared each of these models to choose the best model for prediction using k-folds cross-validation with ten folds, the details of which are covered in the *Evaluation* section. The pros of the models we tested involve the fact that they are easy and computationally cheap to deploy, and the final result of each model is straightforward to interpret. The cons of these models are that they are limited by the variables present in the dataset, as there could be additional factors that could be useful in predicting whether or not an existing health insurance customer is interested in vehicle insurance. With regards to Lasso and Post Lasso, different variables may prove to be useful subjects to the value of the penalty lambda, and so a constraint of these models is choosing the best lambda. The classification tree model is helpful in prediction but makes it difficult to quantify how a change in a predictor variable impacts the dependent variable. Our objective was to choose the best model to predict the likelihood of response and use that probability in a cost-benefit analysis framework to help the firm decide which customers to target to maximize their profit. The formula is as follows:

$$E[Profit \mid X] = P(Interested \mid X) * E[V(i, X) \mid Interested] - Cost(M)$$

X = Customer Features

V(i, X) = Value of providing vehicle insurance to a specific customer

Cost (M) = Cost of marketing (independent of customer-specific features)

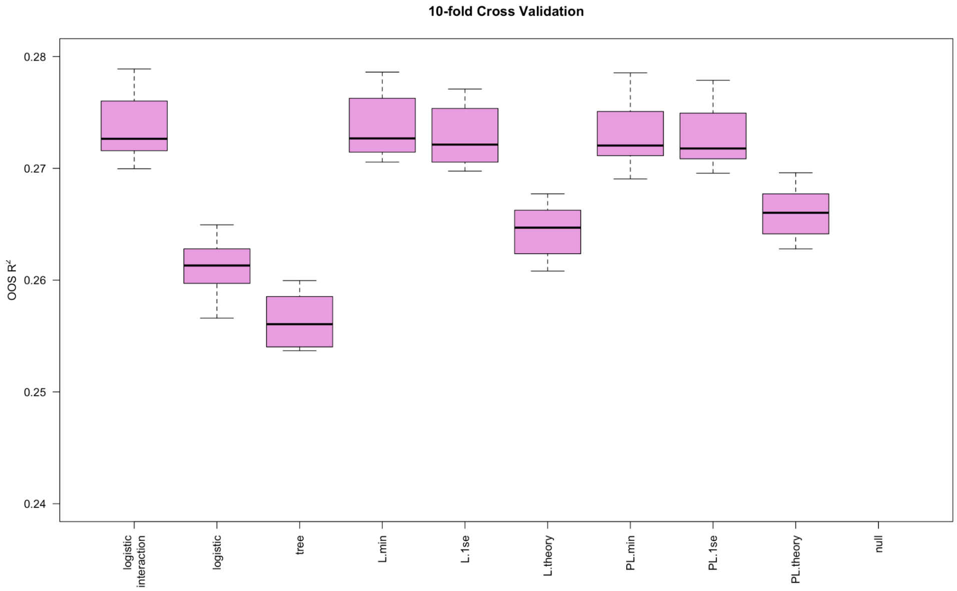
E[Profit | X] = Expected Profit for a given customer

P(Interested | X) = Probability that a specific customer is interested in vehicle insurance

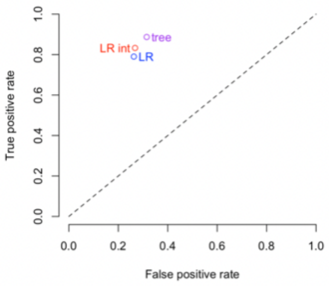
The prediction model selected would allow us to compute a value for the probability of interest for each customer. As for the expected value of providing vehicle insurance to a specific customer, we made assumptions related to how the value of providing vehicle insurance would vary based on certain customer characteristics, such as age and whether or not their vehicles have been damaged in the past. Since the intention is to use this model to identify which customers the company should target, we made assumptions about the cost of marketing. Details of this cost-benefit analysis are covered in the *Deployment* section after selecting the best model for prediction. An alternative modeling approach that we could have taken would have been to predict the likelihood that a customer is interested in vehicle insurance and then deploy a tactic such as a bundling offer of health and vehicle insurance to increase the likelihood of interest for each customer.

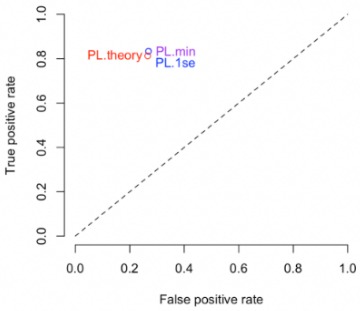
Evaluation

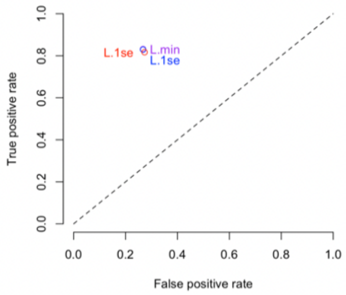
Once we created all our models, we decided to use the 10-fold cross-validation method for comparing out-of-sample metrics. By integrating out-of-sample testing in the cross-validation method, we were able to train and test the model on ten different folds. Each fold allowed us to check the performance of our model when presented with a new but similar dataset. We calculated the out-of-sample R^2 for each of the models and used that to determine the best model, as out-of-sample R^2 measures how well the model fits the test data, by measuring the amount of variation in the dependent variable that is accounted for in the model. We found that our 'LassoMin' model with interaction terms had the highest out-of-sample R^2. The LassoMin takes the smallest lambda (penalty), and this seems to suggest it was the most appropriate model to reduce overfitting. While the Lasso model shrinks the Beta coefficients towards zero for unimportant variables, the LassoMin finds where the minimum mean cross validation error was observed. The LassoMin model narrowed down the interaction terms to 78 variables, with some of the significant interaction terms being *Driver License:Vehicle_age_greater_than_1_year* and *Driver License:Vehicle_age_between_than_1_year_and_2_years*, indicating that people with a driver's license and vehicle age greater than one year are more likely to be interested in the vehicle insurance marketed to them. Note that the graph is zoomed in to be able to compare models, for the original graph with the null model, please refer to the appendix.



In addition to out-of-sample R^2 we calculated and plotted the True Positive Rate (TPR) and False Positive Rate (FPR) for each model to see if they substantially deviated. Our results show that the models have fairly similar TPR/FPR values and so we opted to retain LassoMin as the best model. With regards to the threshold used in the TPR and FPR calculation, we used a threshold of 0.2 instead of 0.5, operating under the assumption that there are asymmetric costs and benefits of marketing to a customer. We assumed the cost of marketing would only be a fraction of the additional premium the firm receives if the customer is interested, and so we lowered the threshold accordingly. In the *Deployment* section, we demonstrate how the firm should apply a different threshold for the probability of interest when deciding whether or not to target a customer, based on the customer’s features.



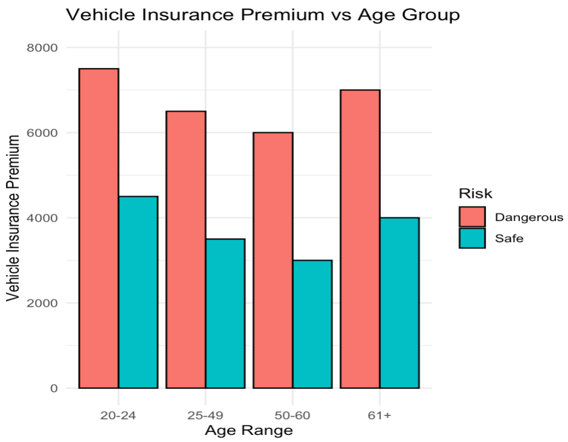




For constant improvement, the business should try to obtain a wider set of customer characteristics. These additional characteristics will provide the models with a larger pool of variables to leverage for prediction. A certain set of variables may prove to be the best possible option, but the current model would not be able to identify this relationship due to the restrictive features. Additionally, the client can provide its existing and potential customers with regular surveys to keep up with the ever-changing trends. With these regular surveys, the business can change the model's parameters as soon as possible and immediately market a more attractive offer to potential customers.

Deployment

Using the LassoMin model with interaction terms, which was the best model based on our cross validation results, the company now has a way to compute the probability of interest in vehicle insurance for all their existing customers. The company can use these calculated probabilities and perform a cost benefit analysis for each customer to determine which customers they should target. Although we did not have the vehicle premiums and marketing cost structures, we made some key assumptions about these values and conducted a cost benefit analysis that the company can reference. We know that vehicle insurance firms tend to charge higher premiums to "risky" individuals, and tend to vary the premiums based on age. We have each customer's age and used "vehicle damage" as a proxy for risk so we were able to generate the following plot, which has our estimates for anticipated vehicle premium each customer would be charged based on their risk and age. The vehicle premium would serve as the value and benefit the company would obtain by providing vehicle insurance to a customer. For more details on how we computed these premiums, please refer to the appendix.



With our anticipated premiums and marketing costs in place, we can plug in these values into our Expected Profit formula (also mentioned in the *Modeling* section):

$$E[Profit \mid X] = P(Interested \mid X) * E[V(i, X) \mid Interested] - Cost(M)$$

Assuming there are no fixed costs, setting the expected profit to 0 and plugging in the expected values and anticipated marketing costs allow us to compute the minimum probability threshold for each customer that the firm should use when deciding whether or not to target a customer. The decision thresholds are outlined in the table below. This shows that for customers that are 25-40 years old and deemed as high risk customers i.e. they have had vehicle damage in the past, the company should market to them even if the probability of interest is as low as 19%. Thus, the company can use our model to predict each customer's interest, and then input their actual anticipated vehicle insurance premiums and realized costs of marketing to determine which customers to target.

Risk Level	**Age**	**Additional Annual Premium (INR)**	**Cost of Marketing (INR)**	**Probability Threshold for Break Even Point**
Safe	20-24	4500	1224	0.27
Safe	25-49	3500	1224	0.35
Safe	50-60	3000	1224	0.41
Safe	61+	4000	1224	0.31
Dangerous	20-24	7500	1224	0.16
Dangerous	25-49	6500	1224	0.19
Dangerous	50-60	6000	1224	0.20
Dangerous	61+	7000	1224	0.17

It is also important to note that this analysis only takes into account the cost of marketing and not the cost of providing vehicle insurance. This is because the cost of providing insurance would require a separate model to estimate the likelihood of the occurrence of an accident. Furthermore, our cost benefit analysis made simplifying assumptions to factor only age and risk level whereas, the company may want to consider additional factors to determine the premium they intend to charge each customer. We recommend that the company use our LassoMin model and cost benefit analysis as a starting point for identifying customers to target, and can then add additional layers of complexity by determining the costs of providing insurance and incorporating the actual vehicle premiums they intend to charge.

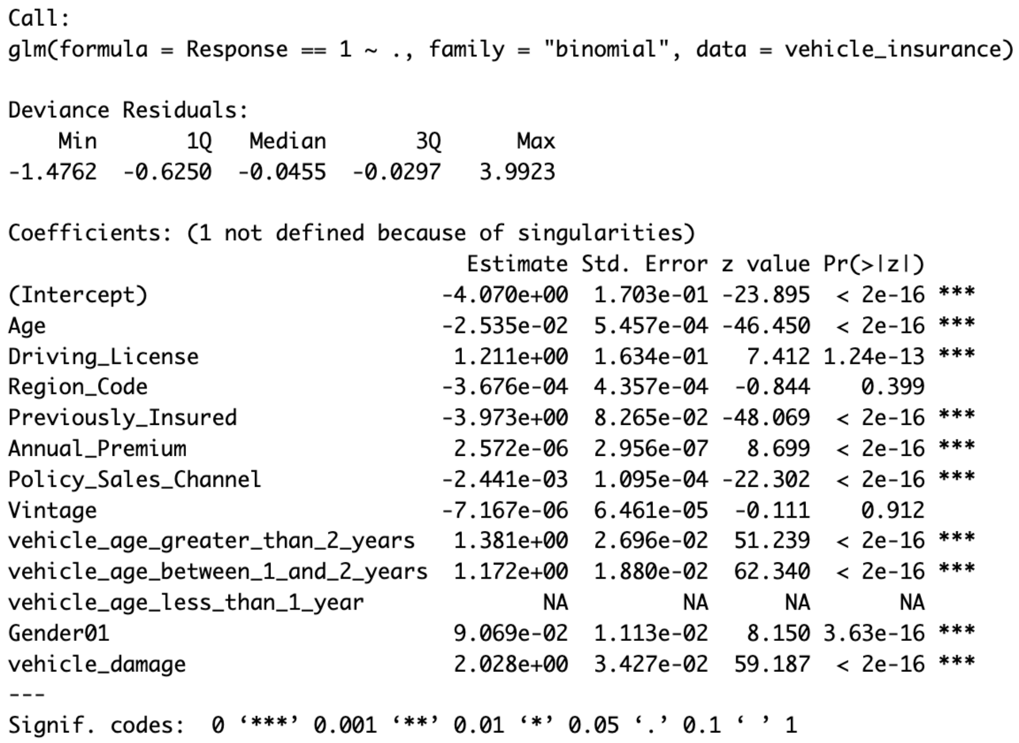
Appendix

Data Source: Dataset: Kumar, Anmol. “Health Insurance Cross Sell Prediction 🏠 🏥.”

Kaggle, 11 Sept. 2020,

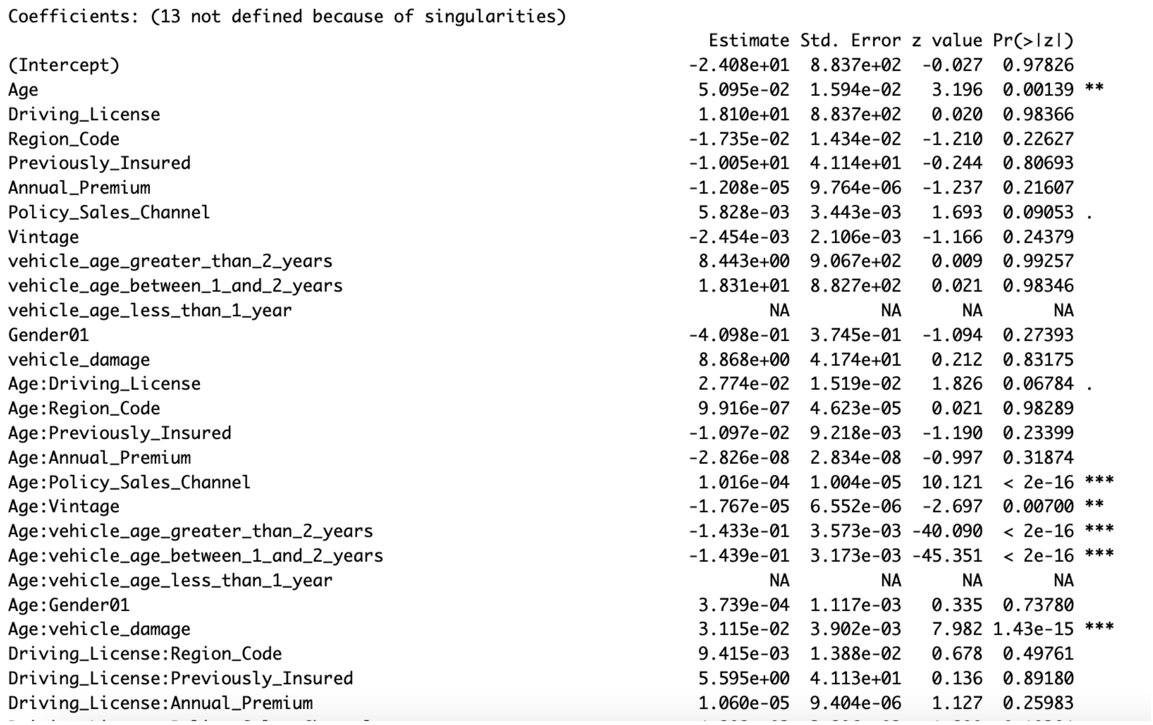
https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction.

Logistic Regression Summary:



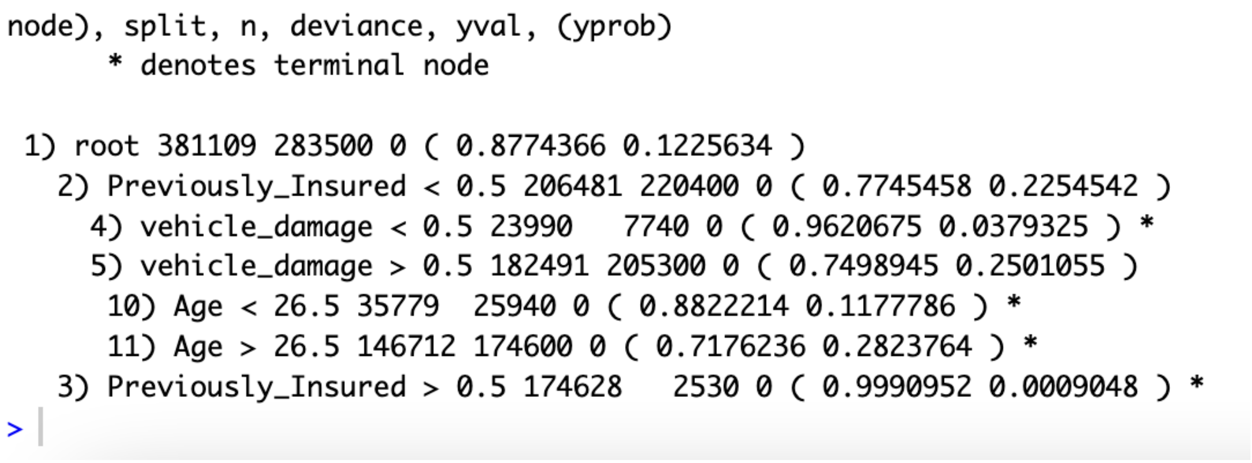
```
Call:
glm(formula = Response == 1 ~ ., family = "binomial", data = vehicle_insurance)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-1.4762  -0.6250  -0.0455  -0.0297   3.9923  

Coefficients: (1 not defined because of singularities)
                                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)                      -4.070e+00  1.703e-01 -23.895  < 2e-16 ***
Age                              -2.535e-02  5.457e-04 -46.450  < 2e-16 ***
Driving_License                   1.211e+00  1.634e-01   7.412 1.24e-13 ***
Region_Code                      -3.676e-04  4.357e-04  -0.844    0.399    
Previously_Insured               -3.973e+00  8.265e-02 -48.069  < 2e-16 ***
Annual_Premium                    2.572e-06  2.956e-07   8.699  < 2e-16 ***
Policy_Sales_Channel             -2.441e-03  1.095e-04 -22.302  < 2e-16 ***
Vintage                          -7.167e-06  6.461e-05  -0.111    0.912    
vehicle_age_greater_than_2_years  1.381e+00  2.696e-02  51.239  < 2e-16 ***
vehicle_age_between_1_and_2_years 1.172e+00  1.880e-02  62.340  < 2e-16 ***
vehicle_age_less_than_1_year             NA         NA      NA       NA    
Gender01                          9.069e-02  1.113e-02   8.150 3.63e-16 ***
vehicle_damage                    2.028e+00  3.427e-02  59.187  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic Interaction Model Summary (just a few variables):



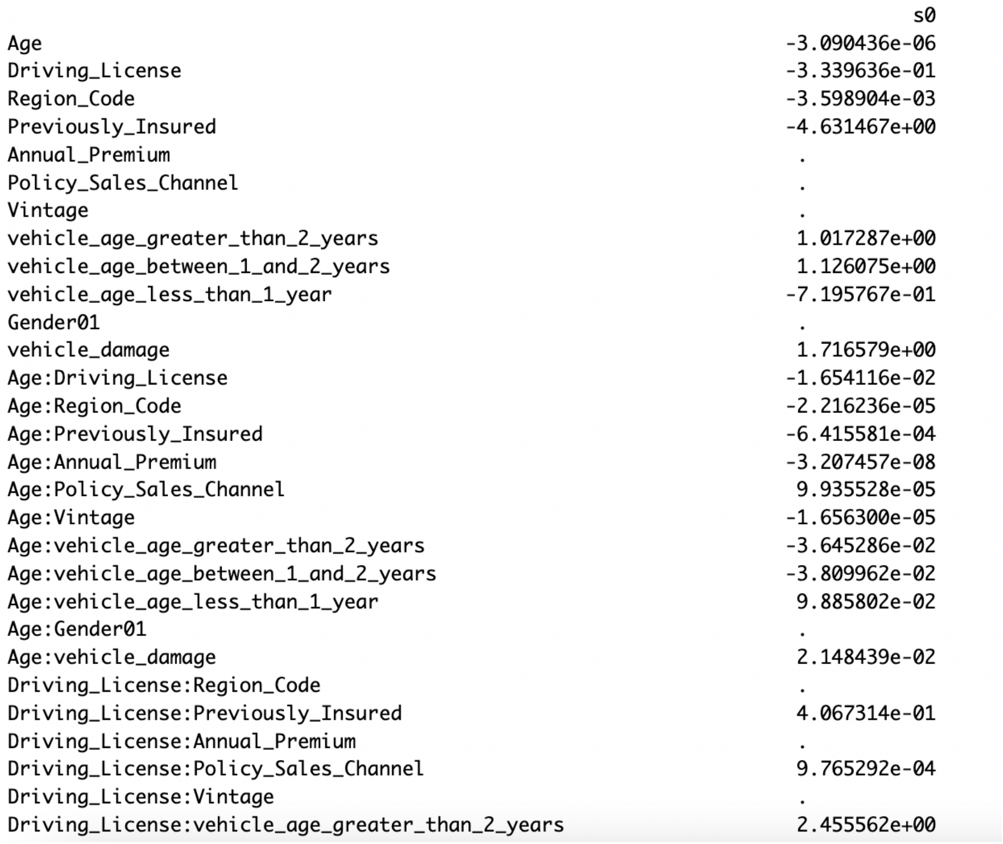
```
Coefficients: (13 not defined because of singularities)
                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       -2.408e+01  8.837e+02  -0.027  0.97826
Age                                                5.095e-02  1.594e-02   3.196  0.00139 **
Driving_License                                    1.810e+01  8.837e+02   0.020  0.98366
Region_Code                                       -1.735e-02  1.434e-02  -1.210  0.22627
Previously_Insured                                -1.005e+01  4.114e+01  -0.244  0.80693
Annual_Premium                                    -1.208e-05  9.764e-06  -1.237  0.21607
Policy_Sales_Channel                               5.828e-03  3.443e-03   1.693  0.09053 .
Vintage                                           -2.454e-03  2.106e-03  -1.166  0.24379
vehicle_age_greater_than_2_years                   8.443e+00  9.067e+02   0.009  0.99257
vehicle_age_between_1_and_2_years                  1.831e+01  8.827e+02   0.021  0.98346
vehicle_age_less_than_1_year                              NA         NA      NA       NA
Gender01                                          -4.098e-01  3.745e-01  -1.094  0.27393
vehicle_damage                                     8.868e+00  4.174e+01   0.212  0.83175
Age:Driving_License                                2.774e-02  1.519e-02   1.826  0.06784 .
Age:Region_Code                                    9.916e-07  4.623e-05   0.021  0.98289
Age:Previously_Insured                            -1.097e-02  9.218e-03  -1.190  0.23399
Age:Annual_Premium                                -2.826e-08  2.834e-08  -0.997  0.31874
Age:Policy_Sales_Channel                           1.016e-04  1.004e-05  10.121  < 2e-16 ***
Age:Vintage                                       -1.767e-05  6.552e-06  -2.697  0.00700 **
Age:vehicle_age_greater_than_2_years              -1.433e-01  3.573e-03 -40.090  < 2e-16 ***
Age:vehicle_age_between_1_and_2_years             -1.439e-01  3.173e-03 -45.351  < 2e-16 ***
Age:vehicle_age_less_than_1_year                          NA         NA      NA       NA
Age:Gender01                                       3.739e-04  1.117e-03   0.335  0.73780
Age:vehicle_damage                                 3.115e-02  3.902e-03   7.982 1.43e-15 ***
Driving_License:Region_Code                        9.415e-03  1.388e-02   0.678  0.49761
Driving_License:Previously_Insured                 5.595e+00  4.113e+01   0.136  0.89180
Driving_License:Annual_Premium                     1.060e-05  9.404e-06   1.127  0.25983
```

Output of Tree Model:



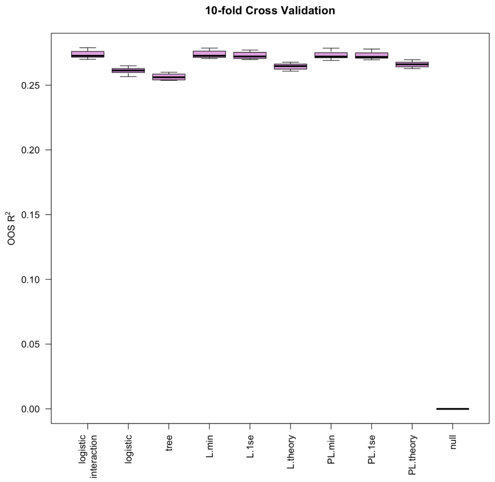
```
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 381109 283500 0 ( 0.8774366 0.1225634 )
   2) Previously_Insured < 0.5 206481 220400 0 ( 0.7745458 0.2254542 )
     4) vehicle_damage < 0.5 23990   7740 0 ( 0.9620675 0.0379325 ) *
     5) vehicle_damage > 0.5 182491 205300 0 ( 0.7498945 0.2501055 )
      10) Age < 26.5 35779  25940 0 ( 0.8822214 0.1177786 ) *
      11) Age > 26.5 146712 174600 0 ( 0.7176236 0.2823764 ) *
   3) Previously_Insured > 0.5 174628   2530 0 ( 0.9990952 0.0009048 ) *
>
```

LassoMin Beta Values (Best Model):



```
                                                          s0
Age                                            -3.090436e-06
Driving_License                                -3.339636e-01
Region_Code                                    -3.598904e-03
Previously_Insured                             -4.631467e+00
Annual_Premium                                  .
Policy_Sales_Channel                            .
Vintage                                         .
vehicle_age_greater_than_2_years                1.017287e+00
vehicle_age_between_1_and_2_years               1.126075e+00
vehicle_age_less_than_1_year                   -7.195767e-01
Gender01                                        .
vehicle_damage                                  1.716579e+00
Age:Driving_License                            -1.654116e-02
Age:Region_Code                                -2.216236e-05
Age:Previously_Insured                         -6.415581e-04
Age:Annual_Premium                             -3.207457e-08
Age:Policy_Sales_Channel                        9.935528e-05
Age:Vintage                                    -1.656300e-05
Age:vehicle_age_greater_than_2_years           -3.645286e-02
Age:vehicle_age_between_1_and_2_years          -3.809962e-02
Age:vehicle_age_less_than_1_year                9.885802e-02
Age:Gender01                                    .
Age:vehicle_damage                              2.148439e-02
Driving_License:Region_Code                     .
Driving_License:Previously_Insured              4.067314e-01
Driving_License:Annual_Premium                  .
Driving_License:Policy_Sales_Channel            9.765292e-04
Driving_License:Vintage                         .
Driving_License:vehicle_age_greater_than_2_years 2.455562e+00
```

OOS R^2 With Null (Zoomed out version)



Cost Benefit Analysis (For Deployment):

While researching, we found that the average cost of vehicle insurance in India is around 5000 rupees. We also found that the cost of vehicle insurance tends to increase with increased risk of the driver and varies with age. We broke down the customers into categories by age and risk, and derived the price for each category by having a different multiple for each category and multiplying the average by that multiple. For example, to get the annual vehicle premium for risky drivers aged 20-25, we used a multiple of 1.5 which resulted in 5000 rupees * 1.5 = 7500 rupees. To get an estimate for safe drivers in that age range, we used a multiple of 0.9 to get 4500 rupees. These multiples are also taken based on reasonable assumptions by looking at the price of vehicle insurance from a variety of vehicle insurance websites. As such, the additional vehicle insurance premium among all categories ranges from 3000 rupees to 7500 rupees.

In order to reach the decision on whether or not to target a customer, we had also to make key assumptions about the associated cost of targeting each customer. The cost of the targeting would remain the same across customer categories, and the detailed cost breakdown follows: 1) We assumed that the average time for targeting a customer is 6 hours, which includes time spent analyzing the customer features, creating marketing material for that customer and reaching out to that customer; 2) the average hourly salary for a marketing associate is 154 rupees; 3) the average cost of company resources for each customer is a flat fee 300 rupees which could include costs of printing documents, mailing letters etc. Putting together, we assume that the total cost for targeting a customer is 1224 rupees.

References for Vehicle Insurance Costs:

"New Car Insurance Price List in India 2022." *PolicyBachat*, https://www.policybachat.com/articles/car-insurance-price-list#:~:text=THIRD%2DPARTY%20INSURANCE%20PREMIUM&text=It%20started%20at%20an%20average,Rs%204000%2D5000%20per%20year.

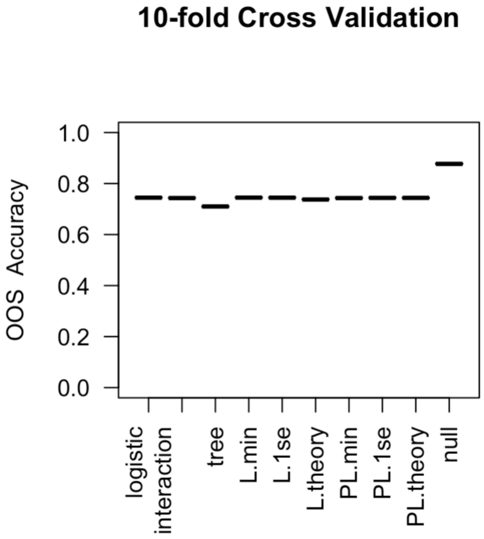
How Much Does Car Insurance Cost on Average? | the Zebra. https://www.thezebra.com/auto-insurance/how-to-shop/average-auto-insurance/.

Deventer, Cate. "Auto Insurance Rates by Age in 2022." *Bankrate*, https://www.bankrate.com/insurance/car/auto-insurance-rates-by-age/#methodology.

"Insurance Associate Salary in India." *AmbitionBox*, https://www.ambitionbox.com/profile/insurance-associate-salary.

Null accuracy:

Since our dataset is imbalanced, the null model in which somebody predicts that nobody is interested, would have an accuracy of 88%. This implies that with no prior knowledge, the firm would be accurate most of the time with the null model, but would miss out on all potential customers. Hence, we opted to not leverage accuracy as our key metric, and emphasized more on out-of-sample R^2 and the TPR and FPR metrics.



Contributions:

Irene Guo: Report Write Up, Visualizations, Powerpoint Slide Preparation

Lucy Meng: Report Write Up, Visualizations, Powerpoint Slide Preparation

Hasan Muhammad: Project Planning, Data Modeling, Model Evaluation, Deployment Analysis, Visualizations, Report Write Up

Manavv Shah: Project Planning, Data Modeling, Model Evaluation, Deployment Analysis, Visualizations, Report Write Up

Tony Qiu: Data Modeling, Model Evaluation, Deployment Analysis, Visualizations, Report Write Up