CS 584: Machine Learning

Spring 2020 Assignment 4

In 2014, Allstate provided the data on Kaggle.com for the Allstate Purchase Prediction Challenge which is open. The data contain transaction history for customers that ended up purchasing a policy. For each Customer ID, you are given their quote history and the coverage options they purchased.

The data is available on the Blackboard as Purchase\_Likelihood.csv.

1. It contains 665,249 observations on 97,009 unique Customer ID.
2. The nominal target variable is **insurance** which has these categories 0, 1, and 2
3. The nominal features are (categories are inside the parentheses):
4. **group\_size**. How many people will be covered under the policy (1, 2, 3 or 4)?
5. **homeowner**. Whether the customer owns a home or not (0 = No, 1 = Yes)?
6. **married\_couple**. Does the customer group contain a married couple (0 = No, 1 = Yes)?

# Question 1 (35 points)

You will build a multinomial logistic model with the following model specifications.

1. Enter the six effects to the model in this sequence:
2. group\_size
3. homeowner
4. married\_couple
5. group\_size \* homeowner
6. group\_size \* married\_couple
7. homeowner \* married\_couple
8. Include the Intercept term in the model
9. The optimization method is Newton
10. The maximum number of iterations is 100
11. The tolerance level is 1e-8.
12. Use the sympy.Matrix().rref() method to identify the non-aliased parameters

Please answer the following questions based on your model.

1. (5 points) List the aliased columns that you found in your model matrix.

* group\_size\_4
* homeowner\_1
* married\_couple\_1
* group\_size\_1 \* homeowner\_1
* group\_size\_2 \* homeowner\_1
* group\_size\_3 \* homeowner\_1
* group\_size\_4 \* homeowner\_0
* group\_size\_4 \* homeowner\_1
* group\_size\_1 \* married\_couple\_1
* group\_size\_2 \* married\_couple\_1
* group\_size\_3 \* married\_couple\_1
* group\_size\_4 \* married\_couple\_0
* group\_size\_4 \* married\_couple\_1
* homeowner\_0 \* married\_couple\_1
* homeowner\_1 \* married\_couple\_0
* homeowner\_1 \* married\_couple\_1

1. (5 points) How many degrees of freedom does your model have?

Degrees of freedom = **2** for Model.

1. (20 points) After entering each model effect, calculate the Deviance test statistic, its degrees of freedom, and its significance value between the current model and the previous model.

List your Deviance test results by the model effects in a table.

| **Step** | **Effect Entered** | **# Free Parameter** | **Log-Likelihood** | **Deviance** | **Degrees of Freedom** | **Significance** |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | Intercept | 2 | -595406.7618844225 | Not Applicable | | |
| 1 | group\_size | 8 | -594912.9735841593 | 987.5766005264595 | 6 | 4.3478703885228946e-210 |
| 2 | homeowner | 10 | -591979.0828339827 | 5867.781500353245 | 2 | 0.0 |
| 3 | married\_couple | 12 | -591936.7938327907 | 84.57800238393247 | 2 | 4.3064572180356084e-19 |
| 4 | group\_size \* homeowner | 18 | -591809.754770109 | 254.07812536344863 | 6 | 5.5121059685664295e-52 |
| 5 | group\_size \* married\_couple | 24 | -591118.4835882675 | 1382.5423636829946 | 6 | 1.4597001210408566e-295 |
| 6 | homeowner \* married\_couple | 26 | -591105.4931771928 | 25.980822149431333 | 2 | 2.28210778553294e-06 |

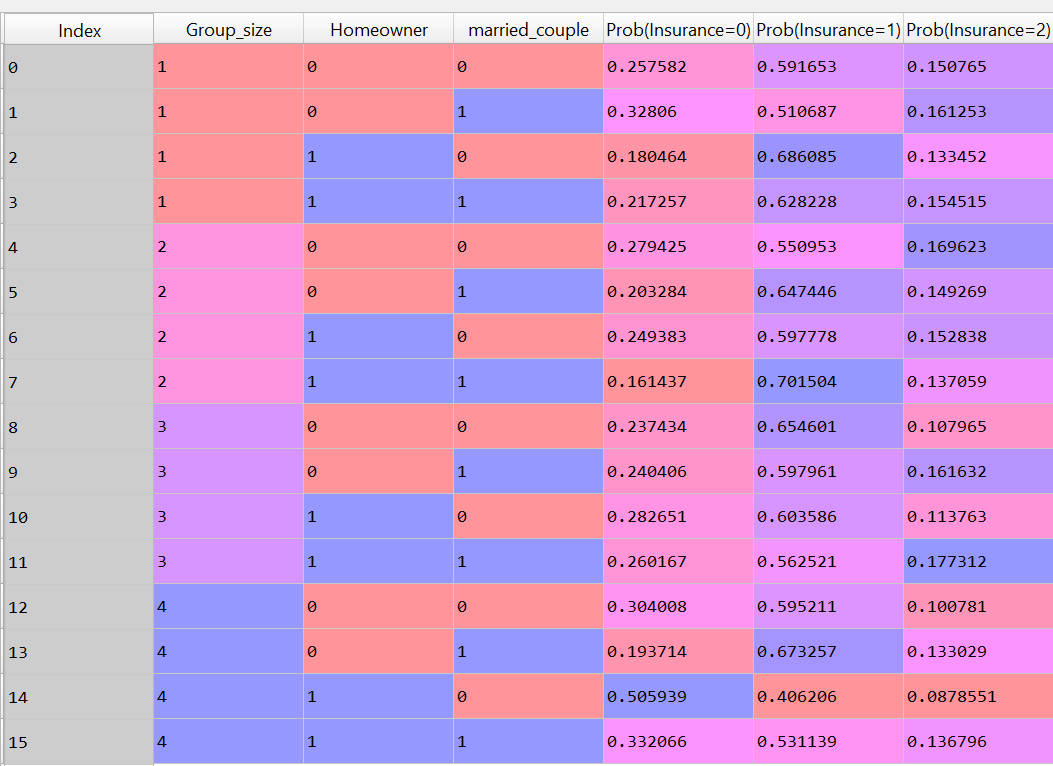
1. (5 points) Calculate the Feature Importance Index as the negative base-10 logarithm of the significance value. List your indices by the model effects.

|  |  |
| --- | --- |
| **Effect Entered** | **Importance** |
| Intercept | Not Applicable |
| group\_size | 209.3617234108572 |
| homeowner | Undefined |
| married\_couple | 18.365879862870976 |
| group\_size \* homeowner | 51.2586824418404 |
| group\_size \* married\_couple | 294.8357363559649 |
| homeowner \* married\_couple | 5.641663847403906 |

# Question 2 (25 points)

Please answer the following questions based on your multinomial logistic model in Question 1.

1. (10 points) For each of the sixteen possible value combinations of the three features, calculate the predicted probabilities for insurance = 0, 1, 2 based on your multinomial logistic model. List your answers in a table with proper labeling.



1. (5 points) Based on your answers in (a), what value combination of group\_size, homeowner, and married\_couple will maximize the odds value Prob(insurance = 1) / Prob(insurance = 0)? What is that maximum odd value?

The following combinations will give will maximize the odds value Prob(insurance = 1) / Prob(insurance = 0).

**group\_size = 2, homeowner = 1, and married\_couple = 1**

The maximum odd value = **4.345370642504374**

1. (5 points) Based on your model, what is the odds ratio for group\_size = 3 versus group\_size = 1, and insurance = 2 versus insurance = 0?  
   (*Hint*: The odds ratio is this odds (Prob(insurance = 2) / Prob(insurance = 0) | group\_size = 3) divided by this odds ((Prob(insurance = 2) / Prob(insurance = 0) | group\_size = 1).)

Taking insurance = 0 as reference target category

Loge((Prob(insurance =2)/Prob(insurance =0) | group\_size = 3) ) –

loge((Prob(insurance =2)/Prob(insurance =0) | group\_size = 1))

= (0.50343 – 0. 546053) + (-0.524617 –(-0.880455))+(0-0) +(0.337205 – 0.902886) + (0 – 0)

= -0.252466

Taking exponent of the previous value: exp(-0.252466) = **0.776882626399578**

1. (5 points) Based on your model, what is the odds ratio for homeowner = 1 versus homeowner = 0, and insurance = 0 versus insurance = 1?

Take insurance =1 as a reference category

(Prob(insurance =0)/Prob(insurance =1) | homeowner = 1) / ((Prob(insurance =0)/Prob(insurance =1) | homeowner = 0)

= Log (Prob(insurance =0)/Prob(insurance =1) | homeowner = 1) - log((Prob(insurance =0)/Prob(insurance =1) | homeowner = 0)

= (0.776052 – 0) +(-1.39531- 0) + (-1.08673 – 0) +( -0.63596 – 0) + (0 – 0) + (0.115368 – 0)

= -2.226584

Taking exponent of the previous value

Exp(-2.226584) = **0.10789637534575976**

# Question 3 (40 points)

You will build a Naïve Bayes model without any smoothing. In other words, the Laplace/Lidstone alpha is zero. Please answer the following questions based on your model.

1. (5 points) Show in a table the frequency counts and the Class Probabilities of the target variable.

|  |  |  |  |
| --- | --- | --- | --- |
| insurance | 0 | 1 | 2 |
| Frequency Count | 143691 | 426067 | 95491 |
| Class Probability | 0.215996 | 0.640462 | 0.143542 |

1. (5 points) Show the crosstabulation table of the target variable by the feature group\_size. The table contains the frequency counts.

|  |  |  |  |
| --- | --- | --- | --- |
| group\_size | insurance | | |
| 0 | 1 | 2 |
| 1 | 115460 | 329552 | 74293 |
| 2 | 25728 | 91065 | 19600 |
| 3 | 2282 | 5069 | 1505 |
| 4 | 221 | 381 | 93 |

1. (5 points) Show the crosstabulation table of the target variable by the feature homeowner. The table contains the frequency counts.

|  |  |  |  |
| --- | --- | --- | --- |
| Homeowner | insurance | | |
| 0 | 1 | 2 |
| 0 | 78659 | 183130 | 46734 |
| 1 | 65032 | 242937 | 48757 |

1. (5 points) Show the crosstabulation table of the target variable by the feature married\_couple. The table contains the frequency counts.

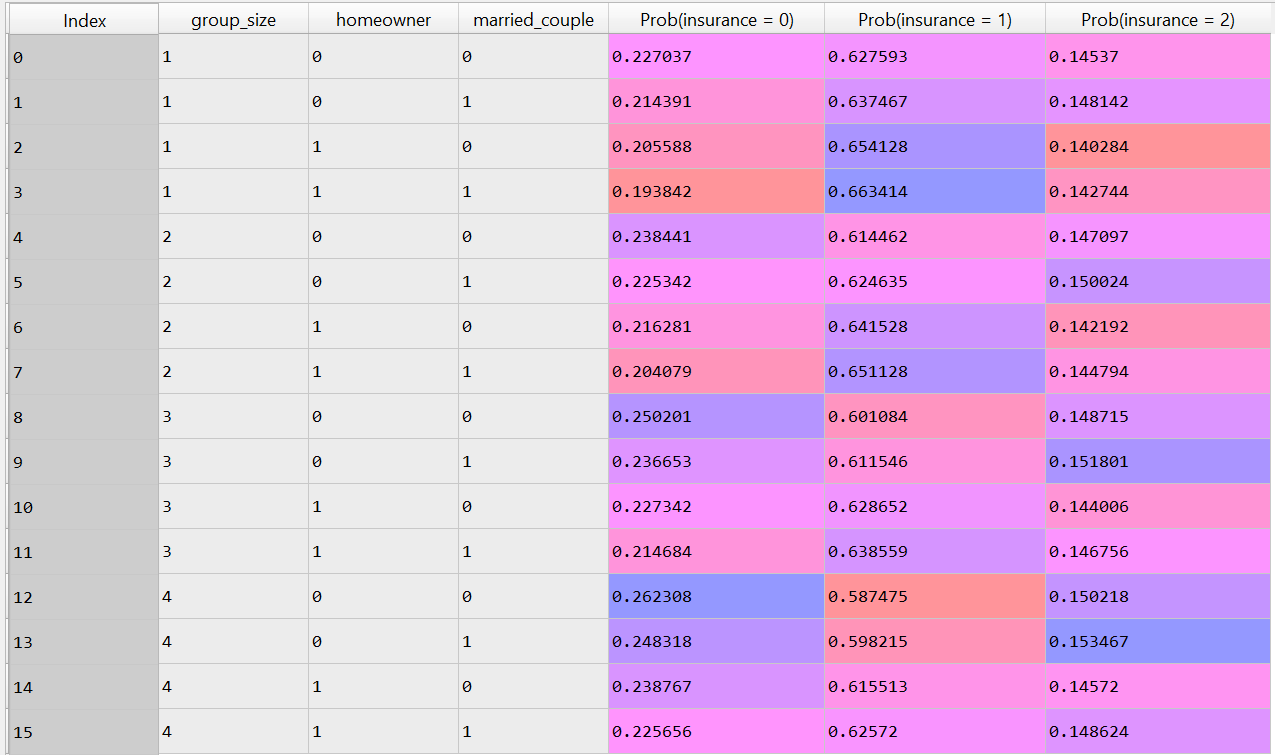
|  |  |  |  |
| --- | --- | --- | --- |
| married\_couple | insurance | | |
| 0 | 1 | 2 |
| 0 | 117110 | 333272 | 75310 |
| 1 | 26581 | 92795 | 20181 |

1. (5 points) Calculate the Cramer’s V statistics for the above three crosstabulations tables. Based on these Cramer’s V statistics, which feature has the largest association with the target insurance?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Feature | Test | Statistic | DF | Significance | Association | Measure |
| group\_size | Chi-  square | 977.276 | 6 | 7.34301e-208 | CramerV | 0.027102 |
| married\_couple | Chi-  square | 699.285 | 2 | 1.41953e-152 | CramerV | 0.0324216 |
| **homeowner** | Chi-  square | 6270.49 | 2 | 0 | CramerV | **0.0970864** |

Based on these Cramer’s V statistics, **homeowner** feature has the largest association with the target insurance.

1. (10 points) For each of the sixteen possible value combinations of the three features, calculate the predicted probabilities for insurance = 0, 1, 2 based on the Naïve Bayes model. List your answers in a table with proper labeling.



1. (5 points) Based on your model, what value combination of group\_size, homeowner, and married\_couple will maximize the odds value Prob(insurance = 1) / Prob(insurance = 0)? What is that maximum odd value?

After the observation of Prob(insurance = 1) / Prob(insurance = 0) for all combinations,

**group\_size = 1**

**Homeowner = 1**

**married\_couple = 1** combination will give maximum odd value for value for Prob(insurance = 1) / Prob(insurance = 0).

The maximize the odds value Prob(insurance = 1) / Prob(insurance = 0) is computed as follow:

Prob(insurance = 0) = 0.193842

Prob(insurance = 1) = 0.663414

The maximum odd value = Prob(insurance = 1) / Prob(insurance = 0)

= 0.663414 / 0.193842

= **3.422441402412735**