CS 584: Machine Learning

Name : Sourav Yadav

AID: A20450418

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):
   a. **group_size**. How many people will be covered under the policy (1, 2, 3 or 4)?
   b. **homeowner**. Whether the customer owns a home or not (0 = No, 1 = Yes)?
   c. **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:
   a. group_size
   b. homeowner
   c. married_couple
   d. group_size * homeowner
   e. group_size * married_couple
   f. homeowner * married_couple
2. Include the Intercept term in the model
3. The optimization method is Newton
4. The maximum number of iterations is 100
5. The tolerance level is 1e-8.
6. Use the sympy.Matrix().rref() method to identify the non-aliased parameters

Please answer the following questions based on your model.

a) (5 points) List the aliased columns that you found in your model matrix.
Answer : List of Aliased Columns for the model :

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

b) (5 points) How many degrees of freedom does your model have?
Degree of Freedom = 2

c) (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**	**-595406.761884422**	**Not Applicable**		
1	group_size	**8**	**-594912.973584159**	**987.576600525993**	**6**	**4.34787038953134E-210**
2	homeowner	**10**	**-591979.082833983**	**5867.78150035347**	**2**	**0**
3	married_couple	**12**	**-591936.793832791**	**84.5780023836996**	**2**	**4.30645721853696E-19**
4	group_size * homeowner	**18**	**-591809.754770109**	**254.078125363449**	**6**	**5.51210596856643E-52**
5	group_size * married_couple	**24**	**-591118.483588268**	**1382.54236368276**	**6**	**1.45970012121037E-295**

Step	Effect Entered	# Free Parameter	Log-Likelihood	Deviance	Degrees of Freedom	Significance
6	homeowner * married_couple	**26**	**-591105.493177193**	**25.980822149897**	**2**	**0.000002821077850016**

d) (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	**209.361723410756**
homeowner	**Infinity**
married_couple	**18.3658798628204**
group_size * homeowner	**51.2586824418404**
group_size * married_couple	**294.835736355914**
homeowner * married_couple	**5.64166384750502**

Question 2 (25 points)

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

a) (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.

group_size	homeOwner	Married_couple	Prob(insurance = 0)	Prob(insurance = 1)	Prob(insurance = 2)
1	0	0	0.257582126	0.591652732	0.150765142
1	0	1	0.328059586	0.510687364	0.16125305
1	1	0	0.180463562	0.686084615	0.133451823
1	1	1	0.217257163	0.628227986	0.154514851
2	0	0	0.279424514	0.550952763	0.169622723
2	0	1	0.203284285	0.647446276	0.149269438
2	1	0	0.249383261	0.597778464	0.152838276
2	1	1	0.161437112	0.701504087	0.137058801
3	0	0	0.237433786	0.654601188	0.107965026
3	0	1	0.240406278	0.597961267	0.161632455
3	1	0	0.282650629	0.603586178	0.113763193

3	1	1	0.260167227	0.562520929	0.177311843
4	0	0	0.304007517	0.595211185	0.100781298
4	0	1	0.193714219	0.673257043	0.133028738
4	1	0	0.505939166	0.406205722	0.087855112
4	1	1	0.332065587	0.531138818	0.136795594

b) (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?

Answer :

Value of Combination

group_size 2

homeowner 1

married_couple 1

Max Odd Value: 4.345370642504375

c) (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).)

Answer :

Odd Ratio: $\frac{\text{(Prob(insurance = 2) / Prob(insurance = 0) | group\_size = 3)}}{\text{(Prob(insurance = 2) / Prob(insurance = 0) | group\_size = 1)}}$

Odds Ratios Group=3 Vs Group=1

```
In [105]: result = pandas.concat([gz_mc_ho,odds],axis=1)
     ...: print(result.iloc[:,1:])
   homeOwner  Married_couple  Odds_ratio
0          0               0    0.776883
1          0               1    1.367816
2          1               0    0.544273
3          1               1    0.958273
```

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

Answer :

Odd Ratio=

$$\frac{(\text{Prob}(\text{insurance} = 0) \,/\, \text{Prob}(\text{insurance} = 1) \mid \text{homeowner} = 1)}{(\text{Prob}(\text{insurance} = 0) \,/\, \text{Prob}(\text{insurance} = 1) \mid \text{homeowner} = 0)}$$

Odd Ratio : HomeOwner 1 Vs HomeOwner 0

```
In [129]: print(result1)
   group_size  homeOwner  Married_couple         0
0           1          1               0  1.655149
1           1          1               1  1.857551
2           2          1               0  1.215691
3           2          1               1  1.364353
4           3          1               0  0.774560
5           3          1               1  0.869278
6           4          1               0  0.410073
7           4          1               1  0.460219
```

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.

a) (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**

b) (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**

c) (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**

d) (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**

e) (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?

Ans:

Cramer's V statistics:

```
Cramers statistics for group_size 0.027102014055820786
Cramers statistics for homeowner 0.09708641964781962
Cramers statistics for married_couple 0.032421645835207 46
the largest association value 0.09708641964781962
```

Homeowner has the highest association with target insurance.

f) (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.

group_size	homeowner	married_couple	Prob(insurance = 0)	Prob(insurance = 1)	Prob(insurance = 2)
1	0	0	0.269722	0.580133	0.150145
1	0	1	0.232789	0.614219	0.152992
1	1	0	0.194038	0.669659	0.136303
1	1	1	0.164935	0.698278	0.136787
2	0	0	0.231143	0.616518	0.152338
2	0	1	0.198016	0.647907	0.154078
2	1	0	0.163628	0.700288	0.136085
2	1	1	0.138274	0.725955	0.135771
3	0	0	0.308219	0.515924	0.175856
3	0	1	0.268311	0.550951	0.180738
3	1	0	0.226972	0.609612	0.163416
3	1	1	0.19437	0.64041	0.165221

group_size	homeowner	married_couple	Prob(insurance = 0)	Prob(insurance = 1)	Prob(insurance = 2)
4	0	0	0.37549	0.48781	0.1367
4	0	1	0.330743	0.527098	0.142158
4	1	0	0.282173	0.588196	0.129631
4	1	1	0.24393	0.623766	0.132304

g) (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?

Ans: The value of combination of group size, homeowner and married couple is (2, 1, 1).

The maximum odd value = 5.250112589270714