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):
    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.

1. **group_size_4**
2. **homeowner_1**
3. **married_couple_1**
4. **group_size_1 * homeowner_1**
5. **group_size_2 * homeowner_1**
6. **group_size_3 * homeowner_1**
7. **group_size_4 * homeowner_0**
8. **group_size_4 * homeowner_1**

9. **group_size_1 * married_couple_1**
10. **group_size_2 * married_couple_1**
11. **group_size_3 * married_couple_1**
12. **group_size_4 * married_couple_0**
13. **group_size_4 * married_couple_1**
14. **homeowner_0 * married_couple_1**
15. **homeowner_1 * married_couple_0**
16. **homeowner_1 * married_couple_1**

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

Degree of Freedom = 26

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	-595406.7618844223	Not Applicable		
1	group_size	8	-594912.9735841593	987.576600	6	4.347870389531338e-210
2	homeowner	10	-591979.0828339825	5867.781500353478	2	0.0
3	married_couple	12	-591936.7938327907	84.578002383699964	2	4.3064572185369587e-19
4	group_size * homeowner	18	-591809.754770109	254.07812536344863	6	5.5121059685664295e-52
5	group_size * married_couple	24	-591118.4835882676	1382.5423636827618	6	1.4597001212103711e-295
6	homeowner * married_couple	26	-591105.4931771928	25.980822149664164	2	2.2821077852672684e-06

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	209.36172341075647
homeowner	**Infinity (because log(0) = negative infinity)**
married_couple	18.365879862820417
group_size * homeowner	51.2586824418404

group_size * married_couple	294.83573635591443
homeowner * married_couple	5.641663847454463

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	Insurance(0)	Insurance(1)	Insurance(2)
0	1	0	0	0.257582	0.591653	0.150765
1	1	0	1	0.328060	0.510687	0.161253
2	1	1	0	0.180464	0.686085	0.133452
3	1	1	1	0.217257	0.628228	0.154515
4	2	0	0	0.279425	0.550953	0.169623
5	2	0	1	0.203284	0.647446	0.149269
6	2	1	0	0.249383	0.597778	0.152838
7	2	1	1	0.161437	0.701504	0.137059
8	3	0	0	0.237434	0.654601	0.107965
9	3	0	1	0.240406	0.597961	0.161632
10	3	1	0	0.282651	0.603586	0.113763
11	3	1	1	0.260167	0.562521	0.177312
12	4	0	0	0.304008	0.595211	0.100781
13	4	0	1	0.193714	0.673257	0.133029
14	4	1	0	0.505939	0.406206	0.087855
15	4	1	1	0.332066	0.531139	0.136796

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?

```
 group_size                 2.000000
homeowner                   1.000000
married_couple              1.000000
Insurance(0)                0.161437
Insurance(1)                0.701504
Insurance(2)                0.137059
odd_value_Insurance(1/0)    4.345371
Name: 7, dtype: float64
```

	group_size	homeowner	married_couple	odd_value_Insurance(1/0)
0	1	0	0	2.296948
1	1	0	1	1.556691
2	1	1	0	3.801790
3	1	1	1	2.891633
4	2	0	0	1.971741
5	2	0	1	3.184930
6	2	1	0	2.397027
7	2	1	1	4.345371
8	3	0	0	2.756984
9	3	0	1	2.487295
10	3	1	0	2.135450
11	3	1	1	2.162151
12	4	0	0	1.957883
13	4	0	1	3.475517
14	4	1	0	0.802875
15	4	1	1	1.599500

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).)

According to the model formulation, the logit Prob(insurance =2)/Prob(insurance =0) for group_size = *r*, homeowner = *s*, and married_couple = *t* is

(Prob(insurance = 2) / Prob(insurance = 0) | group_size = 3) divided by this odds ((Prob(insurance = 2) / Prob(insurance = 0) | group_size = 1)
final equestion of this log_e ((Prob(A=2))/(Prob(A=0)))=μ+ g_r +h_s +m_t + gh_rs +gm_rt + hm_st
r = 1,2,3,4; s = 0,1; t = 0,1
after calculation our final equations will be: **(g_3 - g_1)+(gh_3s - gh_1s)+(gm_3t - gm_1t)**

Homeowner(s)	Married_couple(t)	Odd Ratio:
0	0	0.7768825326496853
0	1	1.3678156191687454
1	0	0.5442730510025144
1	1	0.9582725173582495

Interpretation: If only the group_size of a Insurance is changed from 1 to 3 while keeping all other predictors the same, the odds of Insurance 2 versus Insurance 0 will affect by different factors.

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

group_size	married_couple	Odds Ratio
1	0	0.60417534
1	1	0.53834317
2	0	0.82257763
2	1	0.7329479
3	0	1.29105559
3	1	1.15037954
4	0	2.43859133
4	1	2.17287746

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	115,460	329,552	74,293
2	25,728	91,065	19,600
3	2,282	5,069	1,505
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	78,659	183,130	46,734
1	65,032	242,937	48,757

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	117,110	333,272	75,310
1	26,581	92,795	20,181

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?

Feature	Cramer's V
group_size	0.0271020140558208
homeowner	0.09708641964781962
married_couple	0.03242164583520746

Based on the above Cramer's V statistics, the feature **homeowner** has the largest association with the target variable 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	0.2697219	0.5801334	0.1501447
1	0	1	0.2327892	0.6142186	0.1529922
1	1	0	0.1940379	0.6696590	0.1363031
1	1	1	0.1649350	0.6982780	0.1367869
2	0	0	0.2311433	0.6165185	0.1523382
2	0	1	0.1980156	0.6479068	0.1540776
2	1	0	0.1636275	0.7002878	0.1360847
2	1	1	0.1382742	0.7259550	0.1357709
3	0	0	0.3082194	0.5159242	0.1758564
3	0	1	0.2683111	0.5509509	0.1807380
3	1	0	0.2269718	0.6096118	0.1634164
3	1	1	0.1943695	0.6404098	0.1652207
4	0	0	0.3754904	0.4878101	0.1366995
4	0	1	0.3307434	0.5270983	0.1421583
4	1	0	0.2821727	0.5881965	0.1296309
4	1	1	0.2439303	0.6237660	0.1323037

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?

group_size	homeowner	married_couple	Prob(insurance = 1) / Prob(insurance = 0)
1	0	0	2.1509
1	0	1	2.6385
1	1	0	3.4512
1	1	1	4.2337
2	0	0	2.6673
2	0	1	3.2720
2	1	0	4.2798
2	**1**	**1**	**5.2501**
3	0	0	1.6739
3	0	1	2.0534
3	1	0	2.6858
3	1	1	3.2948
4	0	0	1.2991
4	0	1	1.5937
4	1	0	2.0845
4	1	1	2.5571

```
group_size                    2
homeowner                     1
married_couple                1
Insurance(0)          0.1382742
Insurance(1)          0.7259550
Insurance(2)          0.1357709
odd_value_Insurance(1/0)  5.2501
Name: 3, dtype: object
```

The maximum odd value is **5.2501** that occurs when group_size = 2, homeowner = 1, and married_couple = 1.