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.

The aliased columns that are found in your model matrix are:

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_1
group_size_4 * married_couple_0
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?

Degreee 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	-595406.761884425	Not Applicable		
1	group_size	8	-594912.9735841593	987.5766005264595	6	4.347870388522894 6e-210
2	homeowner	10	-591979.0828339827	5867.781500353245	2	0.0
3	married_couple	12	-591936.7938327906	84.5780023841653	2	4.306457217534288 e-19
4	group_size * homeowner	18	-591809.754770109	254.0781253632158	6	5.512105969198056 e-52
5	group_size * married_couple	24	-591118.4835882675	1382.5423636829946	6	1.459700121040856 6e-295
6	homeowner * married_couple	26	-591105.4931771928	25.980822149431333	2	2.28210778553294e -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	209.3617

homeowner	**inf**
married_couple	**18.3659**
group_size * homeowner	**51.2587**
group_size * married_couple	**294.8357**
homeowner * married_couple	**5.6417**

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	i0	i1	i2
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?

```
Maximum value of insurance=1/insurance=0 is 4.345370642504374
group_size: 2
homeowner: 1
```

```
married_couple: 1
```

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).)

Required Odds Ratio: 1.0249543364157785

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

Required Odds Ratio: 0.6232245044401726

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	381
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.

The crosstabulation table of the target variable by the feature homeowner:

homeowner	0	1
insurance		
0	78659	65032
1	183130	242937
2	46734	48757

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

The crosstabulation table of the target variable by the feature married_couple:

married_couple	0	1
insurance		
0	117110	26581
1	333272	92795
2	75310	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?

```
Cramer's V Value of group size is:  0.027102014055820786
Cramer's V Value of Homeowner is:  0.09708641964781962
Cramer's V Value of Married_couple is:  0.03242164583520746
```

The feature which has largest assocaition with the target insurance is homeowner with 0.09708641964781962

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	0.2270	0.6275	0.1453
1	0	1	0.2143	0.6374	0.1481
1	1	0	0.2055	0.6541	0.1402
1	1	1	0.1938	0.6634	0.1427
2	0	0	0.2384	0.6144	0.1470
2	0	1	0.2253	0.6246	0.1500
2	1	0	0.2162	0.6415	0.1421
2	1	1	0.2040	0.6511	0.1447
3	0	0	0.2502	0.6010	0.1487
3	0	1	0.2366	0.6115	0.1518
3	1	0	0.2273	0.6286	0.1440
3	1	1	0.2146	0.6385	0.1467
4	0	0	0.2623	0.5874	0.1502
4	0	1	0.2483	0.5982	0.1534
4	1	0	0.2387	0.6155	0.1457
4	1	1	0.2256	0.6257	0.1486

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?

Combination:

```
group_size      1
homeowner       1
married_couple  1
```

Maximum odd value(prob(insurance=1)/prob(insurance=0)) = 3.42244