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

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	-595406.7618844225	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

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	209.3617234108572
homeowner	Undefined
married_couple	18.365879862870976

group_size * homeowner	51.2586824418404
group_size * married_couple	294.8357363559649
homeowner * married_couple	294.8357363559649

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.

Index	Group_size	Homeowner	married_couple	Prob(Insurance=0)	Prob(Insurance=1)	Prob(Insurance=2)
0	1	0	0	0.257582	0.591653	0.150765
1	1	0	1	0.32806	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	0.0878551
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 odd value = **4.345370642504374**

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).)

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))

= Parameter of (group_size = 3 | A=2) – Parameter of (group_size = 1 | A=2)

= 0. 50343 – 0. 546053

= -0.042622999999999966

Taking exponent of the previous value: exp(-0.042622999999999966) = **0.9582725907431353**

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

(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 -1.39531*g1 -1.08673*g2-0.63596*g3+0.115368(1-m)
=Exp (Prob(A=0)/Prob(A=1) | homeowner = 1) - log((Prob(A=0)/Prob(A=1) | homeowner = 0)
Here the odds ratio depends on values of group_size and married_couple.
So, g1, g2,g3,m take values of (0 or 1)

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?

	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

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.

Index	group_size	homeowner	married_couple	Prob(insurance = 0)	Prob(insurance = 1)	Prob(insurance = 2)
0	1	0	0	0.227037	0.627593	0.14537
1	1	0	1	0.214391	0.637467	0.148142
2	1	1	0	0.205588	0.654128	0.140284
3	1	1	1	0.193842	0.663414	0.142744
4	2	0	0	0.238441	0.614462	0.147097
5	2	0	1	0.225342	0.624635	0.150024
6	2	1	0	0.216281	0.641528	0.142192
7	2	1	1	0.204079	0.651128	0.144794
8	3	0	0	0.250201	0.601084	0.148715
9	3	0	1	0.236653	0.611546	0.151801
10	3	1	0	0.227342	0.628652	0.144006
11	3	1	1	0.214684	0.638559	0.146756
12	4	0	0	0.262308	0.587475	0.150218
13	4	0	1	0.248318	0.598215	0.153467
14	4	1	0	0.238767	0.615513	0.14572
15	4	1	1	0.225656	0.62572	0.148624

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?

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