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.
Ans:

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?
Ans: 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	-595406.7618844224	Not Applicable		
1	group_size	8	-594912.9735841593	987.5766005262267	6	4.347870389027117e-210
2	homeowner	10	-591979.0828339827	5867.781500353245	2	0.0
3	married_couple	12	-591936.7938327906	84.5780023841653	2	4.306457217534288e-19
4	group_size * homeowner	18	-591809.754770109	254.0781253632158	6	5.512105969198056e-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

Current function value: 0.895013

Iterations 5

MNLogit Regression Results

==========

Dep. Variable: insurance No. Observations: 665249

Model: MNLogit Df Residuals: 665247

Method: MLE Df Model: 0

Date: Wed, 08 Apr 2020 Pseudo R-squ.: 6.440e-11

Time: 20:45:08 Log-Likelihood: -5.9541e+05

converged: True LL-Null: -5.9541e+05

Covariance Type: nonrobust LLR p-value: nan

==========

insurance=1 coef std err z P>|z| [0.025 0.975]

---

insurance 1.0869 0.003 356.296 0.000 1.081 1.093

---

insurance=2 coef std err z P>|z| [0.025 0.975]

---

insurance -0.4086 0.004 -97.874 0.000 -0.417 -0.400

==========

Model Parameter Estimates:

0 1

insurance 1.086931 -0.408633

Model Log-Likelihood Value = -595406.7618844225

Number of Free Parameters = 2

****

Optimization terminated successfully.

Current function value: 0.894271

Iterations 5

Deviance Chi=Square Test

Number of Free Parameters = 8

Model Log-Likelihood Value = -594912.9735841593

Deviance test Statistic = 987.5766005264595

Degreee of Freedom = 6

Significance = 4.3478703885228946e-210

****

Optimization terminated successfully.

Current function value: 0.889861

Iterations 5

Deviance Chi=Square Test

Number of Free Parameters = 10

Model Log-Likelihood Value = -591979.0828339827

Deviance test Statistic = 5867.781500353245

Degreee of Freedom = 2

Significance = 0.0

****

Optimization terminated successfully.

Current function value: 0.889797

Iterations 5

Deviance Chi=Square Test

Number of Free Parameters = 12

Model Log-Likelihood Value = -591936.7938327906

Deviance test Statistic = 84.5780023841653

Degreee of Freedom = 2

Significance = 4.306457217534288e-19

****

Optimization terminated successfully.

Current function value: 0.889606

Iterations 5

Deviance Chi=Square Test

Number of Free Parameters = 18

Model Log-Likelihood Value = -591809.754770109

Deviance test Statistic = 254.0781253632158

Degreee of Freedom = 6

Significance = 5.512105969198056e-52

****

Optimization terminated successfully.

Current function value: 0.888567

Iterations 6

Deviance Chi=Square Test

Number of Free Parameters = 24

Model Log-Likelihood Value = -591118.4835882675

Deviance test Statistic = 1382.5423636829946

Degreee of Freedom = 6

Significance = 1.4597001210408566e-295

Deviance (Statistic, DF, Significance) 1382.5423636829946 6 1.4597001210408566e-295

****

Optimization terminated successfully.

Current function value: 0.888548

Iterations 6

Deviance Chi=Square Test

Number of Free Parameters = 26

Model Log-Likelihood Value = -591105.4931771928

Deviance test Statistic = 25.980822149431333

Degreee of Freedom = 2

Significance = 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.
Ans:

Effect Entered	Importance
Intercept	Not Applicable
group_size	209.36172341080683
homeowner	inf

married_couple	18.36587986292153
group_size * homeowner	51.25868244179064
group_size * married_couple	294.83573635591443
homeowner * married_couple	5.641663847454463

Feature Importance Index for (Intercept + group_size) = 209.36172341080683

Feature Importance Index for (Intercept + group_size + homeowner) = inf

Feature Importance Index for (Intercept + group_size + homeowner + married_couple) = 18.36587986292153

Feature Importance Index for (Intercept + group_size + homeowner + married_couple + group_size * homeowner) = 51.25868244179064

Feature Importance Index for (Intercept + group_size + homeowner + married_couple + group_size * homeowner + group_size * married_couple) = 294.83573635591443

Feature Importance Index for (Intercept + group_size + homeowner + married_couple + group_size * homeowner + group_size * married_couple + 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.
Ans:

group_size	homeowner	married_couple	Prob(insurance = 0)	Prob(insurance = 1)	Prob(insurance = 2)
1	0	0	0.257582	0.591653	0.150765
1	0	1	0.328060	0.510687	0.161253
1	1	0	0.180464	0.686085	0.133452
1	1	1	0.217257	0.628228	0.154515
2	0	0	0.279425	0.550953	0.169623
2	0	1	0.203284	0.647446	0.149269
2	1	0	0.249383	0.597778	0.152838
2	1	1	0.161437	0.701504	0.137059
3	0	0	0.237434	0.654601	0.107965
3	0	1	0.240406	0.597961	0.161632
3	1	0	0.282651	0.603586	0.113763
3	1	1	0.260167	0.562521	0.177312

group_size	homeowner	married_couple	Prob(insurance = 0)	Prob(insurance = 1)	Prob(insurance = 2)
4	0	0	0.304008	0.595211	0.100781
4	0	1	0.193714	0.673257	0.133029
4	1	0	0.505939	0.406206	0.087855
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?

Ans:

	group_size	homeowner	married_couple	odd_value(p_in_1/p_in_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

```
group_size           2.000000
homeowner            1.000000
married_couple       1.000000
0                    0.161437
1                    0.701504
2                    0.137059
odd_value(p_in_1/p_in_0)    4.345371
Name: 7, dtype: float64
```

The maximized odds value of Prob(insurance = 1) / Prob(insurance = 0) is 4.345371 at value combination group_size=2, homeowner=1 & married_couple=1.

The maximum odd value is 4.345371

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).)
Ans: **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?
Ans: **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.
Ans:

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.
Ans:

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.
Ans:

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.
Ans:

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:

	Test	Statistic	DF	Significance	Association	Measure
homeowner	Chi-square	6270.49	2	0	CramerV	0.0970864
married_couple	Chi-square	699.285	2	1.41953e-152	CramerV	0.0324216
group_size	Chi-square	977.276	6	7.34301e-208	CramerV	0.027102

Homeowner has the largest association with the 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.
Ans:

group_size	homeowner	married_couple	Prob(insurance = 0)	Prob(insurance = 1)	Prob(insurance = 2)
1	0	0	0.227037	0.627593	0.145370
1	0	1	0.214391	0.637467	0.148142
1	1	0	0.205588	0.654128	0.140284
1	1	1	0.193842	0.663414	0.142744
2	0	0	0.238441	0.614462	0.147097
2	0	1	0.225342	0.624635	0.150024
2	1	0	0.216281	0.641528	0.142192
2	1	1	0.204079	0.651128	0.144794
3	0	0	0.250201	0.601084	0.148715
3	0	1	0.236653	0.611546	0.151801
3	1	0	0.227342	0.628652	0.144006
3	1	1	0.214684	0.638559	0.146756
4	0	0	0.262308	0.587475	0.150218
4	0	1	0.248318	0.598215	0.153467
4	1	0	0.238767	0.615513	0.145720
4	1	1	0.225656	0.625720	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?

Ans:

	group_size	homeowner	married_couple	odd value(p_in_1/p_in_0)
0	1	0	0	2.764273
1	1	0	1	2.973389
2	1	1	0	3.181743
3	1	1	1	3.422441
4	2	0	0	2.576994
5	2	0	1	2.771943
6	2	1	0	2.966181
7	2	1	1	3.190572
8	3	0	0	2.402403
9	3	0	1	2.584145
10	3	1	0	2.765223
11	3	1	1	2.974412
12	4	0	0	2.239641
13	4	0	1	2.409070
14	4	1	0	2.577880
15	4	1	1	2.772896

```
group_size                1
homeowner                 1
married_couple            1
p_in_0             0.193842
p_in_1             0.663414
p_in_2             0.142744
odd value(p_in_1/p_in_0)    3.42244
Name: 3, dtype: object
```

The maximize odds value of Prob(insurance = 1) / Prob(insurance = 0) is 3.42244 at value combination group_size=1, homeowner=1 & married_couple=1.
The maximum odd value is 3.42244