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_1,
group_size_4* homeowner_0,
homeowner_0* married_couple_1,
homeowner_1* married_couple_1,
homeowner_1* married_couple_0
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

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

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	987.5766005259939	6	4.347870389531338e-210
2	homeowner	10	-591979.0828339525	5867.781500353478	2	0
3	married_couple	12	-591936.7938327907	84.57800238369964	2	4.3064572185369587e-19
4	group_size * homeowner	18	-591909.7547701088	254.07812536368147	6	4.3064572185369587e-19
5	group_size * married_couple	24	-591118.4835882677	1382.5423636822961	6	1.4590012154902334e-295
6	homeowner * married_couple	26	-591105.4931771926	25.980822150129825	2	2.282107784735924e-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
married_couple	18.365879862820417
group_size * homeowner	51.258682441890
group_size * married_couple	294.83573635586
homeowner * married_couple	5.64166384745446

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	0.257582	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
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?

Value of group_size is
2
Value of homeowener is
1
Value of married_couple is
1
Maximum odd Prob(insurance = 1)/Prob(insurance = 0) value is
4.345370642504378

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).)

(Prob(insurance=2)/Prob(insurance=0) | group_size = 3) /
((Prob(insurance=2)/Prob(insurance=0) | group_size = 1) is
0.9582725173582487

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) is
0.4935832413915545

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.

```
The crosstabulation table of the target variable is:

homeowner       0       1     All
A
0           78659   65032  143691
1          183130  242937  426067
2           46734   48757   95491
All        308523  356726  665249
```

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 is:

married_couple       0       1     All
A
0               117110   26581  143691
1               333272   92795  426067
2                75310   20181   95491
All             525692  139557  665249
```

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?

```
The association rank of these three tables are:

                    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

Process finished with exit code 0
```

So the largest one is homeowner.

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	0.26972190083648967	0.5801333993691891	0.15014469979432118
1	0	1	0.23278921851630957	0.6142185578024016	0.15299222368128876
1	1	0	0.19403790475559898	0.6696590048821739	0.1363030903622272
1	1	1	0.164935004743777	0.6982780459509148	0.13678694930530805
2	0	0	0.2311433273249531	0.6165184597447714	0.15233821293027552
2	0	1	0.198015591405003	0.6479067807659843	0.15407762782901277
2	1	0	0.16362752552123652	0.7002878088359464	0.1360846656428170
2	1	1	0.13827417044457968	0.7259549630220522	0.13577086653336812
3	0	0	0.30821939378427693	0.5159241677311622	0.17585643848456095
3	0	1	0.26831105711605896	0.5509508971155715	0.18073804576836952
3	1	0	0.22697183146374494	0.6096117811433283	0.16341638739292683
3	1	1	0.19436951362831584	0.6404097735081213	0.16522071286356266
4	0	0	0.3754903907259939	0.4878101005336526	0.13669950874035344

group_size	homeowner	married_couple	Prob(insurance = 0)	Prob(insurance = 1)	Prob(insurance = 2)
4	0	1	0.3307434441365481	0.527098304946624	0.14215825091682782
4	1	0	0.2821726796029393	0.5881964548622688	0.1296308655347919
4	1	1	0.24393033920041854	0.6237659642682374	0.13230369653134402

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 is
2
married couple is
1
homeowner is
1
the maximum odd is
5.250112589270714
```