CS 584-04: Machine Learning

Autumn 2019 Assignment 4

Question 1 (50 points)

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. It contains 665,249 observations on 97,009 unique Customer ID. You will build a multinomial logistic model with the following specifications.

1. The nominal target variable is **A** which have these categories 0, 1, and 2
2. 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)?
3. Include the Intercept term in the model
4. Enter the five model effects in this order: group_size, homeowner, married_couple, group_size * homeowner, and homeowner * married_couple (No forward or backward selection)
5. The optimization method is Newton
6. The maximum number of iterations is 100
7. The tolerance level is 1e-8.
8. 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 parameters that you found in your model.

Solution:-

group_size_4
homeowner_1
married_couple_1
group_size_1 * home_owner_1
group_size_2 * home_owner_1
group_size_3 * homeowner_1
group_size_4 * homeowner_0
group_size_4 * homeowner_1
homeowner_0 * married_couple_1

homeowner_1 * married_couple_0
homeowner_1 * married_couple_1

b) (5 points) How many degrees of freedom do you have in your model?
Solution:- Degrees of freedom = **20**

c) (10 points) After entering a 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.
Solution:-

Model effect	**Chi-square statistic**	**Degree of freedom**	**Significance**
group_size	987.5766005262267	6	4.347870389027117e-210
homeowner	5867.781500353478	2	0.0
married_couple	84.57800238393247	2	4.3064572180356084e-19
group_size * homeowner	254.07812536344863	6	5.5121059685664295e-52
homeowner * married_couple	70.84227676945738	2	4.13804354793157e-16

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.
Solution:-

Model Effect	**Feature Importance Index**
group_size	209.36172341080683
homeowner	Infinity
married_couple	18.365879862870976
group_size * homeowner	51.2586824418404
homeowner * married_couple	15.383204943219134

e) (10 points) For each of the sixteen possible value combinations of the three features, calculate the predicted probabilities for A = 0, 1, 2 based on the multinomial logistic model. List your answers in a table with proper labelling.

Solution:-

Possibilities	**Pr(A=0)**	**Pr(A=1)**	**Pr(A=2)**
group_size_1_homeowner_0_married_couple_0	0.2596505294674657	0.5891749992224112	0.1511744713101232
group_size_1_homeowner_0_married_couple_1	0.26009172575734996	0.5921057957504566	0.14780247849219338
group_size_1_homeowner_1_married_couple_0	0.18360248712598187	0.6820295479590752	0.13436796491494285
group_size_1_homeowner_1_married_couple_1	0.15402300163900653	0.7099179750480099	0.13605902331298358
group_size_2_homeowner_0_married_couple_0	0.22193613776874224	0.6211051396797336	0.15695872255152415
group_size_2_homeowner_0_married_couple_1	0.2223208689711648	0.6242161616144875	0.15346296941396015
group_size_2_homeowner_1_married_couple_0	0.20250954944289226	0.6597726841350491	0.13771776642205874
group_size_2_homeowner_1_married_couple_1	0.17055155513697823	0.6894495093334505	0.13999893552957146
group_size_3_homeowner_0_married_couple_0	0.23957008219988554	0.6046159207044135	0.15581399709597948
group_size_3_homeowner_0_married_couple_1	0.23999174875571322	0.6076604706960655	0.15234778054822107
group_size_3_homeowner_1_married_couple_0	0.30113985065740032	0.5312967796225498	0.16756336972004698
group_size_3_homeowner_1_married_couple_1	0.25901737310942163	0.5670166436209104	0.17396598326966792
group_size_4_homeowner_0_married_couple_0	0.19448468744537793	0.6696859053777223	0.13582940717689840
group_size_4_homeowner_0_married_couple_1	0.19469210100526804	0.6725920901301082	0.13271580886462372
group_size_4_homeowner_1_married_couple_0	0.38771909269820077	0.4849744467497125	0.12730646055207973
group_size_4_homeowner_1_married_couple_1	0.33917171209639435	0.5264040860788882	0.13442420182471734

f) (5 points) Based on your model, what values of group_size, homeowner, and married_couple will maximize the odds value Prob(A=1) / Prob(A = 0)? What is that maximum odd value?

Solution:- group_size = **1**, homeowner = **1**, married_couple = **1**

Maximum odd value = **4.609168549460486**

g) (5 points) Based on your model, what is the odds ratio f or group_size = 3 versus group_size = 1, and A = 2 versus A = 0? Mathematically, the odds ratio is (Prob(A=2)/Prob(A=0) | group_size = 3) / ((Prob(A=2)/Prob(A=0) | group_size = 1).

Solution:-
(Prob(A=2)/Prob(A=0) | group_size = 3) / ((Prob(A=2)/Prob(A=0) | group_size = 1) =>
log((Prob(A=2)/Prob(A=0) | group_size = 3) / ((Prob(A=2)/Prob(A=0) | group_size = 1))=
log(Prob(A=2)/Prob(A=0) | group_size = 3) – log(Prob(A=2)/Prob(A=0) | group_size = 1)

➔ (group_size_3 – group_size_1)+ (group_size_3*homeowner_0 - group_size_1*homeowner_0)(1-h)
➔ **(-0.274022 + 0.384741)*(1-h)**

(Prob(A=2)/Prob(A=0) | group_size = 3) / ((Prob(A=2)/Prob(A=0) | group_size = 1)
⇨exp (0.527471 – 0.801493) + (-0.5987 + 0.983441)(1-h)
⇨**exp ((-0.274022 + 0.384741)*(1-h))**
Here the odds ratio depends on the values of group_size and homeowner. So h take values of (0 or 1) where h is homeowner.

h) (5 points) Based on your model, what is the odds ratio for homeowner = 1 versus homeowner = 0, and A = 0 versus A = 1? Mathematically, the odds ratio is (Prob(A=0)/Prob(A=1) | homeowner = 1) / ((Prob(A=0)/Prob(A=1) | homeowner = 0).

Solution:- (Prob(A=0)/Prob(A=1) | homeowner = 1) / ((Prob(A=0)/Prob(A=1) | homeowner = 0) =>
log(Prob(A=0)/Prob(A=1) | homeowner = 1) – log(Prob(A=0)/Prob(A=1) | homeowner = 0) =
(0.800157 – 1.505554 * g1 – 1.164638 * g2 – 0.654639 * g3 + 0.212483 *(1-m).

(Prob(A=0)/Prob(A=1) | homeowner = 1) / ((Prob(A=0)/Prob(A=1) | homeowner = 0)
➔ **exp ((0.800157 – 1.505554 * g1 – 1.164638 * g2 – 0.654639 * g3 + 0.212483 *(1-m)).**

Here the odds ratio depends on the values of group_size and married_couple.
So g1 , g2, , g3 , m take values of (0 or 1) where g1, g2, g3 is group_size_1, group_size_2, group_size_3 respectively. And m is married_couple.

Question 2 (50 points)

You are asked to build a Naïve Bayes model using the same Purchase_Likelihood.csv. The model specifications are:

1. No smoothing is needed. Therefore, the Laplace/Lidstone alpha is zero
2. The nominal target variable is **A** which have 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)?

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.
Solution:-

	A	Count	Proportion
0	0	143691	0.215996
1	1	426067	0.640462
2	2	95491	0.143542

b) (5 points) Show the crosstabulation table of the target variable by the feature group_size. The table contains the frequency counts.
Solution:-

group_size				
A	**1**	**2**	**3**	**4**
0	115460	25728	2282	221
1	329552	91065	5069	381
2	74293	19600	1505	93

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

Solution:-

homeowner		
A	**0**	**1**
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.

Solution:-

married_couple		
A	**0**	**1**
0	117110	26581
1	333272	92795
2	75310	20181

e) (10 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 A?

Solution:-

Cramer's V Value for group_size is: **0.027102014055820786**
Cramer's V Value for homeowner is: **0.09708641964781962**
Cramer's V Value for married_couple is: **0.03242164583520746**
"homeowner" has the largest association with the target A.

f) (5 points) Based on the assumptions of the Naïve Bayes model, express the joint probability Prob(A = a, group_size = g, homeowner = h, married_couple = m) as a product of the appropriate probabilities.

Solution:-

Prob(A=a, group_size=g, homeowner=h, married_couple=m) =
Prob(group_size=g, homeowner=h, married_couple=m |A=a)*Prob(A=a) =
Prob(group_size=g | A=a)*Prob(homeowner=h | A=a)*Prob(married_couple=m | A=a)*Prob(A=a)

g) (10 points) For each of the sixteen possible value combinations of the three features, calculate the predicted probabilities for A = 0, 1, 2 based on the Naïve Bayes model. List your answers in a table with proper labelling.

Solution:-

Possibilities	**Pr(A=0)**	**Pr(A=1)**	**Pr(A=2)**
group_size_1_homeowner_0_married_couple_0	0.26972190083648967	0.5801333993691891	0.15014469979432118
group_size_1_homeowner_0_married_couple_1	0.23278921851630957	0.6142185578024016	0.15299222368128876
group_size_1_homeowner_1_married_couple_0	0.19403790475559898	0.6696590048821739	0.136303090 3622272
group_size_1_homeowner_1_married_couple_1	0.164935004743777	0.6982780459509148	0.13678694930530805
group_size_2_homeowner_0_married_couple_0	0.2311433273249531	0.6165184597447714	0.15233821293027552
group_size_2_homeowner_0_married_couple_1	0.198015591405003	0.6479067807659843	0.15407762782901277
group_size_2_homeowner_1_married_couple_0	0.16362752552123652	0.7002878088359464	0.13608466564281702
group_size_2_homeowner_1_married_couple_1	0.13827417044457968	0.7259549630220522	0.13577086653336812
group_size_3_homeowner_0_married_couple_0	0.30821939378427693	0.5159241677311622	0.17585643848456095
group_size_3_homeowner_0_married_couple_1	0.26831105711605896	0.5509508971155715	0.18073804576836952
group_size_3_homeowner_1_married_couple_0	0.22697183146374494	0.6096117811433283	0.16341638739292683
group_size_3_homeowner_1_married_couple_1	0.19436951362831584	0.6404097735081213	0.16522071286356266
group_size_4_homeowner_0_married_couple_0	0.37549039007259939	0.4878101005336526	0.13669950874035344
group_size_4_homeowner_0_married_couple_1	0.33074344411365481	0.5270983040946624	0.14215825091682782
group_size_4_homeowner_1_married_couple_0	0.28217267960293393	0.5881964548622688	0.12963086553347919
group_size_4_homeowner_1_married_couple_1	0.24393033920041854	0.6237659642682374	0.13230369653134402

h) (5 points) Based on your model, what values of group_size, homeowner, and married_couple will maximize the odds value Prob(A=1) / Prob(A = 0)? What is that maximum odd value?

Solution:-

The maximum value occurs when group_size = **2**, homeowner = **1**, married_couple = **1**.
The maximum value is: **5.250112589270714**