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.

Explanation:

- Aliasing (or redundancy) indicates that a column in the design matrix is linearly dependent on other columns. In other words, that column can be expressed as a linear combination of the other columns. The aliased parameters found in the model are shown in the following table.

Aliased Parameter
Group_size_1
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
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?

Explanation:

The degree of freedom for this model is **2**.

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.

Explanation:

- Deviance test statistics: The deviance is basically a measure of how much unexplained variation there is in our logistic regression model – the higher the value the less accurate the model.
- Degree of freedom: Each of a number of independently variable factors affecting the range of states in which a system may exist, in particular any of the directions in which independent motion can occur.
- Significance value: Is the probability of rejecting the null hypothesis when it is true.

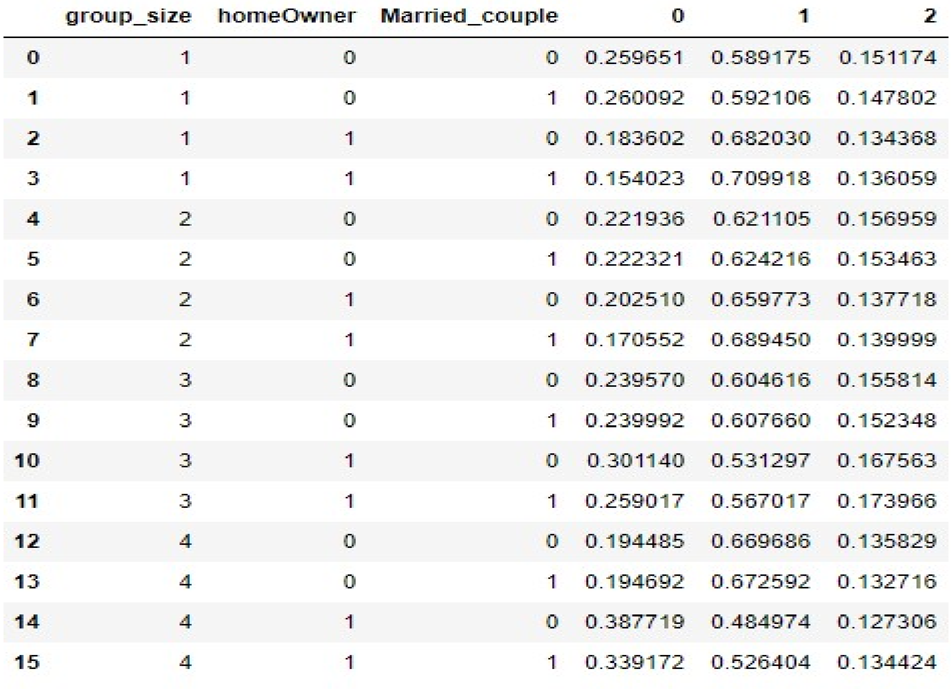
Variables	**Deviance Test Statistics(Chi2)**	**Degree of Freedom**	**Significance Value**
Intercept + group_size	987.5766005262267	6	4.347870389027117e-210
Intercept + group_size + homeowner	5867.781500353245	2	0
Intercept + group_size + homeowner + married_couple	84.5780023841653	2	4.306457217534288e-19
Intercept + group_size + homeowner + married_couple + group_size*homeowner	254.0781253632158	6	5.512105969198056e-52
Intercept + group_size + homeowner + married_couple + group_size*homeowner + homeowner*married_couple	70.84227677015588	2	4.13804354648637e-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.

Explanation:

Variables	**Feature Importance**
Intercept + group_size	209.36172341080683
Intercept + group_size + homeowner	Infinity
Intercept + group_size + homeowner + married_couple	18.36587986292153
Intercept + group_size + homeowner + married_couple + GroupSize*Homeowner	51.25868244179064
Intercept + group_size + homeowner + married_couple + group_size*homeowner + homeowner*married_couple	15.38320494337081

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.



	group_size	homeOwner	Married_couple	0	1	2
0	1	0	0	0.259651	0.589175	0.151174
1	1	0	1	0.260092	0.592106	0.147802
2	1	1	0	0.183602	0.682030	0.134368
3	1	1	1	0.154023	0.709918	0.136059
4	2	0	0	0.221936	0.621105	0.156959
5	2	0	1	0.222321	0.624216	0.153463
6	2	1	0	0.202510	0.659773	0.137718
7	2	1	1	0.170552	0.689450	0.139999
8	3	0	0	0.239570	0.604616	0.155814
9	3	0	1	0.239992	0.607660	0.152348
10	3	1	0	0.301140	0.531297	0.167563
11	3	1	1	0.259017	0.567017	0.173966
12	4	0	0	0.194485	0.669686	0.135829
13	4	0	1	0.194692	0.672592	0.132716
14	4	1	0	0.387719	0.484974	0.127306
15	4	1	1	0.339172	0.526404	0.134424

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?

Explanation:

Calculation for the combination of [group_size,homeowner,married_couple] = [1,1,1], we get the maximum odd value for Prob(A=1) / Prob(A = 0). Maximum odd value that is observed is 4.609169.

g) (5 points) Based on your model, what is the odds ratio for 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).

Explanation:

Taking A=0 as reference target category = Log e((Prob(A=2)/Prob(A=0) | group_size = 3)) – log e((Prob(A=2)/Prob(A=0) | group_size = 1))

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

= 0.527471 - 0.801493

= -0.274022

Taking exponent of the previous value: exp(-0.274022) = 0.76031534813

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).

Explanation:

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)

Exp (Prob(A=0)/Prob(A=1) | homeowner = 1) - log((Prob(A=0)/Prob(A=1) | homeowner = 0)

Question 2 (50 points)

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

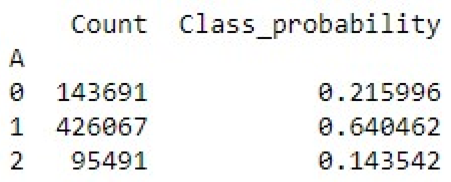
1. No smoothing is needed. Therefore, the Laplace/Lidstone alpha is zero
2. The nominal target variable is **A** which have these categories 0, 1, and 2
3. The nominal features are (categories are inside the parentheses):
   a. **group_size**. How many people will be covered under the policy (1, 2, 3 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.

Explanation:

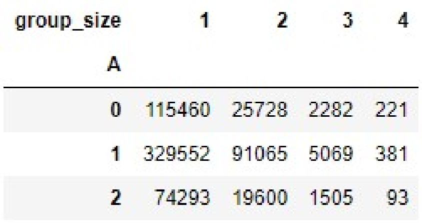
Class Probabilities is calculated for each class such that (no. of values of each class/ total no. of values)



```
    Count  Class_probability
A
0  143691           0.215996
1  426067           0.640462
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.

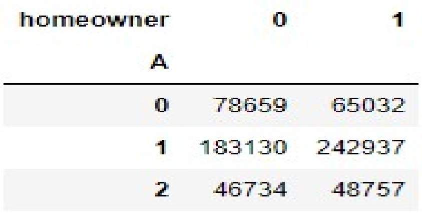
Explanation:



group_size	1	2	3	4
A				
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.

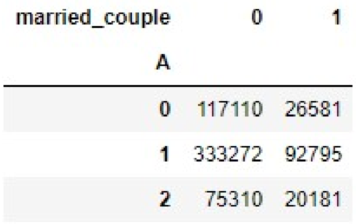
Explanation:



homeowner	0	1
A		
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.

Explanation:



married_couple	0	1
A		
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?

Explanation:

We use Cramers V to find association between target and predictors. Cramers V indicate strength of association of two nominal variable. Higher value shows higher association between the specified variables.

Feature	**Cramers's V**
HomeOwner	0.0970864
Married_couple	0.0324216
Group_size	0.027102

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.

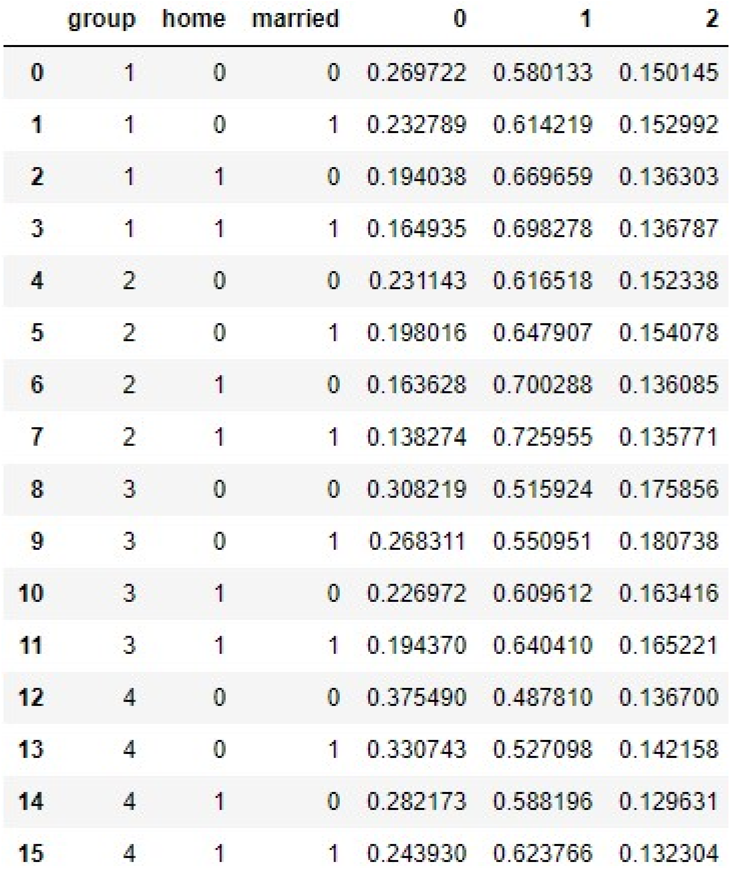
Explanation:

Joint Probability Prob(A = a, group_size = g, homeowner = h, married_couple = m) is given as:

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

Explanation:



	group	home	married	0	1	2
0	1	0	0	0.269722	0.580133	0.150145
1	1	0	1	0.232789	0.614219	0.152992
2	1	1	0	0.194038	0.669659	0.136303
3	1	1	1	0.164935	0.698278	0.136787
4	2	0	0	0.231143	0.616518	0.152338
5	2	0	1	0.198016	0.647907	0.154078
6	2	1	0	0.163628	0.700288	0.136085
7	2	1	1	0.138274	0.725955	0.135771
8	3	0	0	0.308219	0.515924	0.175856
9	3	0	1	0.268311	0.550951	0.180738
10	3	1	0	0.226972	0.609612	0.163416
11	3	1	1	0.194370	0.640410	0.165221
12	4	0	0	0.375490	0.487810	0.136700
13	4	0	1	0.330743	0.527098	0.142158
14	4	1	0	0.282173	0.588196	0.129631
15	4	1	1	0.243930	0.623766	0.132304

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?

Explanation:

The maximum value = [group_size , homeowner , married_couple] = [2,1,1]

The maximum odds value for Prob(A=1) / Prob(A = 0) is 5.250113