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):
3. **group\_size**. How many people will be covered under the policy (1, 2, 3 or 4)?
4. **homeowner**. Whether the customer owns a home or not (0 = No, 1 = Yes)?
5. **married\_couple**. Does the customer group contain a married couple (0 = No, 1 = Yes)?
6. Include the Intercept term in the model
7. 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)
8. The optimization method is Newton
9. The maximum number of iterations is 100
10. The tolerance level is 1e-8.
11. Use the sympy.Matrix().rref() method to identify the non-aliased parameters

Please answer the following questions based on your model.

1. (5 points) List the aliased parameters that you found in your model.

These are the non-aliased parameters:

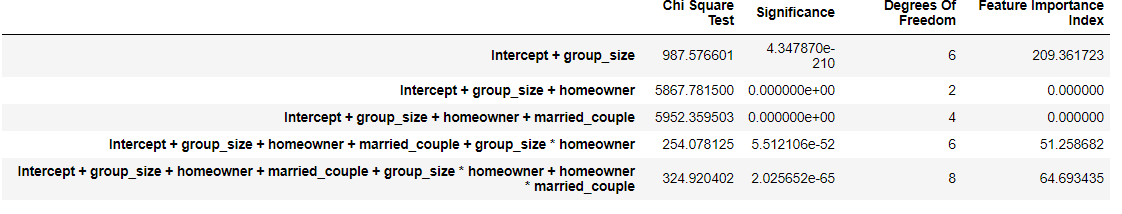
(0, 1, 2, 3, 5, 7, 9, 11, 13)

**So all the columns except these can be listed as the aliased paramet-ers**

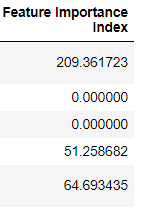
1. (5 points) How many degrees of freedom do you have in your model?

Degree of freedom obtained = 8

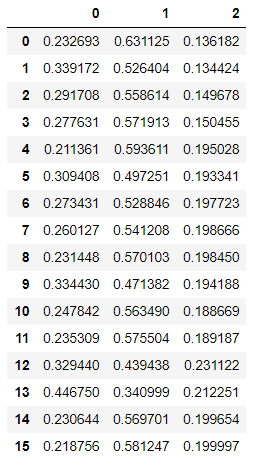
1. (10 points) After entering a model effect, calculate the Deviance test statistic, its degrees of freedom, and its significance value between the current model and the previous model. List your Deviance test results by the model effects in a table.



1. (5 points) Calculate the Feature Importance Index as the negative base-10 logarithm of the significance value. List your indices by the model effects.



1. (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.



1. (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?

**The maximum odd value is 4.610**

1. (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).
2. (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).

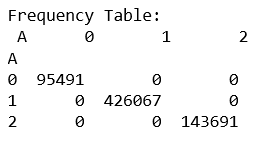
# Question 2 (50 points)

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

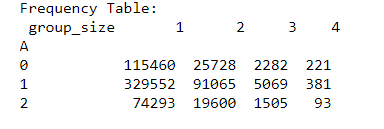
1. No smoothing is needed. Therefore, the Laplace/Lidstone alpha is zero
2. The nominal target variable is **A** which have these categories 0, 1, and 2
3. The nominal features are (categories are inside the parentheses):
4. **group\_size**. How many people will be covered under the policy (1, 2, 3 or 4)?
5. **homeowner**. Whether the customer owns a home or not (0 = No, 1 = Yes)?
6. **married\_couple**. Does the customer group contain a married couple (0 = No, 1 = Yes)?

Please answer the following questions based on your model.

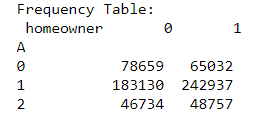
1. (5 points) Show in a table the frequency counts and the Class Probabilities of the target variable.



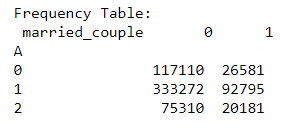
1. (5 points) Show the crosstabulation table of the target variable by the feature group\_size. The table contains the frequency counts.



1. (5 points) Show the crosstabulation table of the target variable by the feature homeowner. The table contains the frequency counts.



1. (5 points) Show the crosstabulation table of the target variable by the feature married\_couple. The table contains the frequency counts.



1. (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?

Observed Count:

1 74293 115460 329552

col\_0

0 0 1 0

1 0 0 1

2 1 0 0

Column Total:

col\_0

0 1

1 1

2 1

dtype: int64

Row Total:

1

74293 1

115460 1

329552 1

dtype: int64

Overall Total:

3

Expected Count:

[[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]]

For Feature 1 1.0

Observed Count:

col\_0 0 1 2

2

19600 0 0 1

25728 1 0 0

91065 0 1 0

Column Total:

2

19600 1

25728 1

91065 1

dtype: int64

Row Total:

col\_0

0 1

1 1

2 1

dtype: int64

Overall Total:

3

Expected Count:

[[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]]

For Feature 2 1.0

Observed Count:

col\_0 0 1 2

3

1505 0 0 1

2282 1 0 0

5069 0 1 0

Column Total:

3

1505 1

2282 1

5069 1

dtype: int64

Row Total:

col\_0

0 1

1 1

2 1

dtype: int64

Overall Total:

3

Expected Count:

[[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]]

For Feature 3 1.0

Observed Count:

col\_0 0 1 2

4

93 0 0 1

221 1 0 0

381 0 1 0

Column Total:

4

93 1

221 1

381 1

dtype: int64

Row Total:

col\_0

0 1

1 1

2 1

dtype: int64

Overall Total:

3

Expected Count:

[[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]]

For Feature 4 1.0

Observed Count:

col\_0 0 1 2

0

46734 0 0 1

78659 1 0 0

183130 0 1 0

Column Total:

0

46734 1

78659 1

183130 1

dtype: int64

Row Total:

col\_0

0 1

1 1

2 1

dtype: int64

Overall Total:

3

Expected Count:

[[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]]

For Feature 0 1.0

Observed Count:

1 48757 65032 242937

col\_0

0 0 1 0

1 0 0 1

2 1 0 0

Column Total:

col\_0

0 1

1 1

2 1

dtype: int64

Row Total:

1

48757 1

65032 1

242937 1

dtype: int64

Overall Total:

3

Expected Count:

[[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]]

For Feature 1 1.0

Observed Count:

col\_0 0 1 2

0

75310 0 0 1

117110 1 0 0

333272 0 1 0

Column Total:

0

75310 1

117110 1

333272 1

dtype: int64

Row Total:

col\_0

0 1

1 1

2 1

dtype: int64

Overall Total:

3

Expected Count:

[[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]]

For Feature 0 1.0

Observed Count:

1 20181 26581 92795

col\_0

0 0 1 0

1 0 0 1

2 1 0 0

Column Total:

col\_0

0 1

1 1

2 1

dtype: int64

Row Total:

1

20181 1

26581 1

92795 1

dtype: int64

Overall Total:

3

Expected Count:

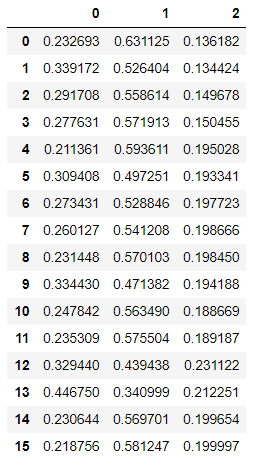
[[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]

[0.33333333 0.33333333 0.33333333]]

For Feature 1 1.0

1. (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.
2. (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.



1. (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?

Maximum odd value is 5.250 and the combination is (2,1,1)