CS 484: Introduction to Machine Learning

Autumn 2020 Assignment 4

# Question 1 and 2

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** that has 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)*?

# Question 1 (35 points)

You will train a multinomial logistic model with the following model specifications.

1. Use all 665,249 observations for training the model
2. Enter the six effects to the model in the **following** order:
3. group\_size
4. homeowner
5. married\_couple
6. group\_size \* homeowner
7. group\_size \* married\_couple
8. homeowner \* married\_couple
9. Include the Intercept term in the model
10. Use the SWEEP Operator method to identify the non-aliased parameters
11. The optimization method is Newton
12. The maximum number of iterations is 100
13. The tolerance level is 1e-8.

Please answer the following questions based on your model.

1. (5 points) List the **aliased** columns that you have identified 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\_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']**

1. (5 points) How many degrees of freedom does your model have?

**24**

1. (20 points) After entering each effect into the current model, 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** | **-595406.761884** | **Not Applicable** | | |
| **1** | **group\_size** | **8** | **-594912.973584** | **987.577** | **6** | **4.34787e-210** |
| **2** | **Homeowner** | **10** | **-591979.082834** | **6855.36** | **8** | **0** |
| **3** | **married\_couple** | **12** | **-591936.793833** | **6939.94** | **10** | **0** |
| **4** | **group\_size \* homeowner** | **18** | **-591809.754770** | **7194.01** | **16** | **0** |
| **5** | **group\_size \* married\_couple** | **24** | **-591118.483588** | **8576.56** | **22** | **0** |
| **6** | **homeowner \* married\_couple** | **26** | **-591105.493177** | **8602.54** | **24** | **0** |

1. (5 points) Calculate the Feature Importance Index as the negative base-10 logarithm of the significance value. If the significance value is zero, then assign Infinity to Importance. List your indices by the model effects.

| **Effect Entered** | **Importance** |
| --- | --- |
| **Intercept** | **Not Applicable** |
| **group\_size** | **209.3617234108572** |
| **Homeowner** | **inf** |
| **married\_couple** | **Inf** |
| **group\_size \* homeowner** | **Inf** |
| **group\_size \* married\_couple** | **Inf** |
| **homeowner \* married\_couple** | **Inf** |

# Question 2 (40 points)

You will train a Naïve Bayes model without any smoothing using all the observations in the **Purchase\_Likelihood.csv**. In other words, the Laplace/Lidstone alpha is zero. 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.

|  |  |  |  |
| --- | --- | --- | --- |
| **insurance** | **0** | **1** | **2** |
| **Frequency Count** | **143691** | **426067** | **95491** |
| **Class Probability** | **0.216** | **0.64** | **0.144** |

1. (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** |

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

|  |  |  |  |
| --- | --- | --- | --- |
| **homeowner** | **insurance** | | |
| **0** | **1** | **2** |
| **0** | **78659** | **183130** | **46734** |
| **1** | **65034** | **242937** | **48757** |

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Married\_couple** | **insurance** | | |
| **0** | **1** | **2** |
| **0** | **117110** | **333272** | **75310** |
| **1** | **26581** | **92795** | **20181** |

1. (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?

|  |  |
| --- | --- |
| **Feature** | **Cramer’s V** |
| **group\_size** | **0.0271** |
| **homeowner** | **0.097** |
| **married\_couple** | **0.0324** |

**Homeowner has the largest Cramer’s V, thus homeowner has the largest association with the target insurance.**

1. (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 that includes features group\_size, homeowner, and married\_couple. 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** | **0.227** | **0.627** | **0.145** |
| **1** | **0** | **1** | **0.214** | **0.637** | **0.148** |
| **1** | **1** | **0** | **0.205** | **0.654** | **0.140** |
| **1** | **1** | **1** | **0.193** | **0.663** | **0.142** |
| **2** | **0** | **0** | **0.238** | **0.614** | **0.147** |
| **2** | **0** | **1** | **0.225** | **0.624** | **0.15** |
| **2** | **1** | **0** | **0.216** | **0.641** | **0.142** |
| **2** | **1** | **1** | **0.204** | **0.651** | **0.144** |
| **3** | **0** | **0** | **0.250** | **0.601** | **0.148** |
| **3** | **0** | **1** | **0.236** | **0.611** | **0.151** |
| **3** | **1** | **0** | **0.227** | **0.628** | **0.144** |
| **3** | **1** | **1** | **0.214** | **0.638** | **0.146** |
| **4** | **0** | **0** | **0.262** | **0.587** | **0.150** |
| **4** | **0** | **1** | **0.248** | **0.598** | **0.153** |
| **4** | **1** | **0** | **0.238** | **0.615** | **0.145** |
| **4** | **1** | **1** | **0.225** | **0.625** | **0.148** |

1. (5 points) Based on your model, what value combination of group\_size, homeowner, and married\_couple will yield the maximum odds value Prob(insurance = 1) / Prob(insurance = 0)? What is that maximum odd value?

**The max is 3.422 from 1,1,1 (group\_size = 1, homeowner = 1, married\_couple = 1).**

# Question 3 (10 points)

You will calculate the Eta-squared statistic to measure the association between the interval target MPG\_Highway and the categorical feature DriveTrain which has three categories: *All*, *Front*, and *Rear*. Instead of the original training data, you are given the following table of summary statistics.

|  |  |  |  |
| --- | --- | --- | --- |
| **DriveTrain** | **Count** | **Mean** | **Corrected Sum of Squares** |
| All | 92 | 22.4673913043478 | 1574.9021739130400 |
| Front | 226 | 29.5044247787611 | 7794.4955752212400 |
| Rear | 110 | 25.0363636363636 | 983.8545454545450 |
| **Total** | **428** | **26.8434579439252** | **14074.5116822429000** |

**SSW = 1574.9 + 7794.5 + 983.9 = 10353.3**

**SST = 14074.511**

**SSG = SST – SSW = 14074.511 – 10353.3 = 3721.211**

**Eta-squared = SSG / SST = 3721.211 / 14074.511 = 0.264**

# Question 4 (15 points)

You live in the San Francisco Bay area where earthquakes are not uncommon. Your house has a security alarm system against burglary, and it can be set off occasionally by an earthquake. Historically, there is a 6% chance that your house will be burglarized and there is a 2% chance that an earthquake will occur in your area. You can assume that the occurrences of burglary and earthquake are statistically independent. Based on your experience, your alarm will sound if the following events have occurred.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Earthquake** | True | True | False | False |
| **Burglary** | True | False | True | False |
| **Probability that Alarm will sound** | 0.99 | 0.15 | 0.95 | 0.0001 |

Please calculate this quantity Prob(Burglary = True and Earthquake = False | Alarm Sounded = False), i.e., the conditional probability that your house has been burglarized but no earthquake has occurred provided the alarm has not sounded.

**Prob(Burglary = True and Earthquake = False | Alarm Sounded = False)**

**= Prob(Burglary = True and Earthquake = False and Alarm Sounded = False) / Prob(Alarm Sounded = False)**

**Independent so:**

**= Prob(Burglary = True) \* Prob(Earthquake = False) \* Prob(Alarm Sounded = False) / Prob(Alarm Sounded = false)**

**= 0.06 \* 0.98 \* 0.05 / 0.05 = 0.0588??**

**Hmmm or**

* **Pr(A|B) = ( Pr(B|A) Pr(A) ) / ( Pr(B|A) P(A) + Pr(B|~A) P(~A) )**

**(((0.06 \* 0.98 \* 0.05)/(0.06\*0.98))\*(0.06\*0.98))/((((0.06 \* 0.98 \* 0.05)/(0.06\*0.98))\*(0.06\*0.98))+(((0.94\*0.02\*0.85)/(0.94\*0.02))\*(0.94\*0.02)))**

**= 0.155**