Tania Soutonglang

CS 484: Introduction to Machine Learning

Fall Semester 2022 Assignment 4

# Question 1 (50 Points)

In 2014, Allstate provided the data on Kaggle.com for the Allstate Purchase Prediction Challenge. The data contains transaction history for customers that ended up purchasing a policy. For each Customer ID, we know the quote history and the purchased coverage options.

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 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)*?

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 Counts | 143,691 | 426,067 | 95,491 |
| Class Probabilities | 0.21599582 | 0.64046244 | 0.14354174 |

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** | **1** | **2** | **3** | **4** | **Total** |
| **0** | 115,460 | 25,728 | 2,282 | 221 | 143,691 |
| **1** | 329,552 | 91,065 | 5,069 | 381 | 426,067 |
| **2** | 74,293 | 19,600 | 1,505 | 93 | 95,491 |
| **Total** | 519,305 | 136,393 | 8,856 | 695 | 665,249 |

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** | **Total** |
| **0** | 78,659 | 65,032 | 143,691 |
| **1** | 183,130 | 242,937 | 426,067 |
| **2** | 46,734 | 48,757 | 95,491 |
| **Total** | 308,523 | 356,726 | 665,249 |

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** | **Total** |
| **0** | 117,110 | 26,581 | 143,691 |
| **1** | 333,272 | 92,795 | 426,067 |
| **2** | 75,310 | 20,181 | 95,491 |
| **Total** | 525,692 | 139,557 | 665,249 |

1. (5 points) Calculate the Cramer’s V statistics for the above three crosstabulations tables. Based on the Cramer’s V statistics, which feature has the strongest association with the target insurance?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Expected Values (E)** | **group\_size** | | | |
| **insurance** | **1** | **2** | **3** | **4** |
| **0** |  |  |  |  |
| **1** |  |  |  |  |
| **2** |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Squared Distance** | **group\_size** | | | |
| **Insurance** | **1** | **2** | **3** | **4** |
| **0** | 96.6338 | 472.8460 | 71.2364 | 33.4698 |
| **1** | 27.8476 | 157.6003 | 64.0930 | 9.2369 |
| **2** | 0.8314 | 0.0245 | 42.9984 | 0.4583 |

|  |  |  |
| --- | --- | --- |
| **Expected Values** | **homeowner** | |
| **insurance** | **0** | **1** |
| **0** | 66,639.6769 | 77,051.3231 |
| **1** | 197,597.3944 | 228,469.6056 |
| **2** | 44,285.9287 | 51,205.0713 |

|  |  |  |
| --- | --- | --- |
| **Squared Distance** | **Homeowner** | |
| **insurance** | **0** | **1** |
| **0** | 2,167.8396 | 1,874.9078 |
| **1** | 1,059.2523 | 916.1197 |
| **2** | 135.3264 | 117.0402 |

|  |  |  |
| --- | --- | --- |
| **Expected Values** | **married\_couple** | |
| **Insurance** | **0** | **1** |
| **0** | 113,547.2720 | 30,143.7280 |
| **1** | 336,685.9828 | 89,381.0172 |
| **2** | 75,458.7452 | 20,032.2548 |

|  |  |  |
| --- | --- | --- |
| **Squared Distance** | **married\_couple** | |
| **Insurance** | **0** | **1** |
| **0** | 111.7863 | 421.0836 |
| **1** | 34.6177 | 130.3999 |
| **2** | 0.2932 | 1.1045 |

|  |  |
| --- | --- |
| **Feature** | **Cramer’s V** |
| group\_size | 0.0271 |
| **homeowner** | **0.0971** |
| married\_couple | 0.0324 |

From the calculations, homeowner has the strongest association to the target.

1. (15 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** | **Insurance=0** | **Insurance=1** | **Insurance=2** |
| 1 | 0 | 0 | 0.269722 | 0.580133 | 0.150145 |
| 1 | 0 | 1 | 0.232789 | 0.614219 | 0.152992 |
| 1 | 1 | 0 | 0.194038 | 0.669659 | 0.136303 |
| 1 | 1 | 1 | 0.164935 | 0.698278 | 0.136787 |
| 2 | 0 | 0 | 0.231143 | 0.616518 | 0.152338 |
| 2 | 0 | 1 | 0.198016 | 0.647907 | 0.154078 |
| 2 | 1 | 0 | 0.163628 | 0.700288 | 0.136085 |
| 2 | 1 | 1 | 0.138274 | 0.725955 | 0.135771 |
| 3 | 0 | 0 | 0.308219 | 0.515924 | 0.175856 |
| 3 | 0 | 1 | 0.268311 | 0.550951 | 0.180738 |
| 3 | 1 | 0 | 0.226972 | 0.609612 | 0.163416 |
| 3 | 1 | 1 | 0.19437 | 0.64041 | 0.165221 |
| 4 | 0 | 0 | 0.37549 | 0.48781 | 0.1367 |
| 4 | 0 | 1 | 0.330743 | 0.527098 | 0.142158 |
| 4 | 1 | 0 | 0.282173 | 0.588196 | 0.129631 |
| 4 | 1 | 1 | 0.24393 | 0.623766 | 0.132304 |

1. (10 points) Based on your model, determine the value combination of group\_size, homeowner, and married\_couple that will yield the maximum value for this odds Prob(insurance = 1) / Prob(insurance = 2)? What is that maximum odds value?

|  |  |  |  |
| --- | --- | --- | --- |
| **group\_size** | **homeowner** | **married\_couple** | **Prob(insurance = 1)/Prob(insurance = 2)** |
| 1 | 0 | 0 | 3.8638 |
| 1 | 0 | 1 | 4.0147 |
| 1 | 1 | 0 | 4.9130 |
| 1 | 1 | 1 | 5.1049 |
| 2 | 0 | 0 | 4.0470 |
| 2 | 0 | 1 | 4.2051 |
| 2 | 1 | 0 | 5.1460 |
| **2** | **1** | **1** | **5.3469** |
| 3 | 0 | 0 | 2.9338 |
| 3 | 0 | 1 | 3.0483 |
| 3 | 1 | 0 | 3.7304 |
| 3 | 1 | 1 | 3.8761 |
| 4 | 0 | 0 | 3.5685 |
| 4 | 0 | 1 | 3.7078 |
| 4 | 1 | 0 | 4.5375 |
| 4 | 1 | 1 | 4.7146 |

Group size: 2, Homeowner: Yes, Married Couple: Yes will yield a maximum value of **5.3469.**

# Question 2 (50 points)

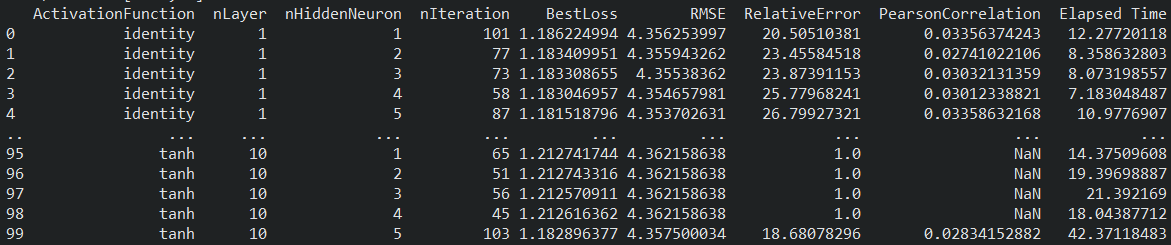
The **Homeowner\_Claim\_History.xlsx** contains the claim history of 27,513 homeowner policies. The following table describes the eleven columns in the HOCLAIMDATA sheet.

| **Name** | **Description** | **Categories** |
| --- | --- | --- |
| policy | Policy Identifier |  |
| exposure | Duration a Policy is Exposed to Risk Measured in Portion of a Year |  |
| num\_claims | Number of Claims in a Year |  |
| amt\_claims | Total Claim Amount in a Year |  |
| f\_primary\_age\_tier | Age Tier of Primary Insured | < 21, 21 - 27, 28 - 37, 38 - 60, > 60 |
| f\_primary\_gender | Gender of Primary Insured | Female, Male |
| f\_marital | Marital Status of Primary Insured | Not Married, Married, Un-Married |
| f\_residence\_location | Location of Residence Property | Urban, Suburban, Rural |
| f\_fire\_alarm\_type | Fire Alarm Type | None, Standalone, Alarm Service |
| f\_mile\_fire\_station | Distance to Nearest Fire Station | < 1 mile, 1 - 5 miles, 6 - 10 miles, > 10 miles |
| f\_aoi\_tier | Amount of Insurance Tier | < 100K, 100K - 350K, 351K - 600K, 601K - 1M, > 1M |

In insurance ratemaking, the ratio of Number of Claims in a Year divided by the Duration a Policy is Exposed to Risk is called the Frequency. In other words, Frequency is the average number of claims per year. If a policy does not file any claims in a year, then its Frequency is zero.

1. (20 points). Train a Multi-Layer Perceptron neural network. The target variable is Frequency. The predictors are the seven categorical predictors. Perform a grid search to select the most desired network structure. The maximum number of iterations is 10000. The random seed is 31010. Try the Hyperbolic Tangent and the Identity activation functions, the number of layers from 1 to 10, and the common number of neurons per layer from 1 to 5. Please be reminded that the answer key of this question is prepared using sklearn version 1.1.2.

Show your grid search results in a table. The table should contain (1) the activation function type, (2) the number of layers, (3) the common number of neurons per layer, (4) the number of iterations performed (*n\_iter\_* attribute), (5) the best loss value (*best\_loss\_* attribute), (6) the root mean squared error, (7) the relative error, (8) the Pearson correlation, and (9) the elapsed time in seconds.



1. (10 points) Among the networks that converged, please recommend the most desired network structure which yields the lowest Root Mean Squared Error. In the case of ties, choose the network with a fewer total number of neurons.

A picture containing graphical user interface

Description automatically generated

1 layer and 3 neurons is the most desirable network structure with a Root Mean Squared Error of 4.3457.

1. (10 points) Assess the final model goodness-of-fit using (1) Root Mean Squared Error, (2) Relative Error, and (3) Pearson Correlation. What are the values of these metrics?

The final model has an RMSE of 4.3457, a relative error of 34.0872, and a Pearson Correlation of 0.0571.

1. (10 points) Identify any poorly predicted observations. First, plot the predicted versus the observed Frequency. Second, together in a single chart frame, plot the Simple Residuals and the Pearson Residuals versus the observed Frequency. Label the axes of these two charts accordingly. To receive full credits, generate your charts with proper dimensions (e.g., length and width) and resolution (e.g., dpi). The Simple Residual equals to the observed Frequency minus the predicted Frequency. The Pearson Residual equals the Simple Residual divided by the square root of the predicted Frequency.