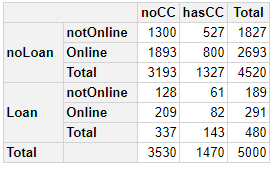
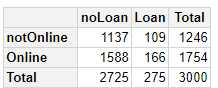
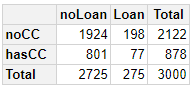
Chapter 8 Problems

1. Problem 8.1
   1. The table below shows the pivot tables



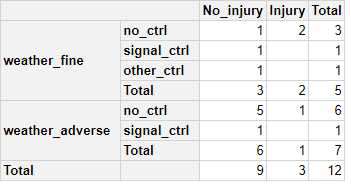
* 1. 82/882 = 0.093
  2. The two pivot tables below for online, loan, and credit card data on the training dataset.





* 1. Calculations
     1. P(CC | Loan) = 0.28
     2. P (Online | Loan) = 0.60
     3. P(Loan) = 0.092
     4. P(CC | noLoan) = 0.72
     5. P( Online | noLoan) = 0.58
     6. P(noLoan) = 0.908
  2. Using naïve Bayes, P( Loan | CC, Online) = 0.039
  3. The value is not that close. They are both low, under 10%, but the naïve bayes method is a little further from the actual to be considered accurate.
  4. An entry that corresponds to CC=1 and online =1 is the 13th entry. The model predicts a personal loan 9% of the time. This corresponds directly to the value in B, but not the value in e. This means I must have calculated something wrong.

1. Problem 8.2
   1. 49.7% of the time there is an injury.
   2. First 12 records questions
      1. Pivot table below



* + 1. P(Injury) = 25%,

P(Injury | no\_ctrl) = 33%,

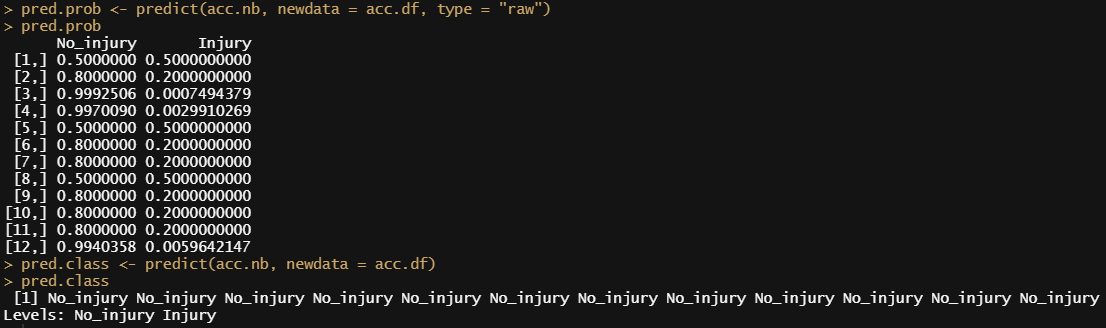
P(Injury | signal\_ctrl) = 0 %

P(Injury| other\_ctrl) = 0 %

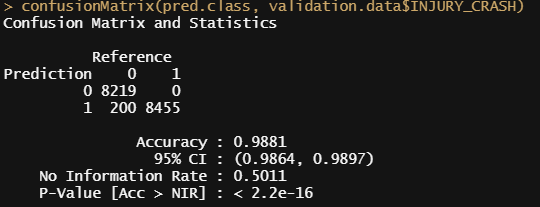
P(Injury | weather\_fine) = 40%

P (Injury | weather\_adverse) = 14%

* + 1. Using a cutoff of 50%, no record would be predicted to have an injury.
    2. The naïve bayes calculation would be 0% because the probability of P(TRAF\_CON\_R =1 | Injury) is 0, and that is in the numerator. Therefor, it will always be 0%.
    3. The naïve bayes model classifies all records as no injury. Below are the results for both class prediction and probability prediction.



* 1. Entire dataset questions
     1. The only predictors that should not be included are number of vehicles involved, speed limit, and number of injuries based on the fact that they are not categorical, or at least purely. Maximum severe injury and fatalities should not be included because they are only associated with injuries.
     2. Confusion Matrix is shown below.



* + 1. Overall error is 1.1%
    2. Unsure.
    3. I decided not to use speed limit in the initial model. Going back and adding it in, the answer is 0 because there were not enough samples.