Problem 1:

Data Processing: First I separated that data into that we had labels for and those that we did not. The data for which we had labels, I then split randomly into training and test data with a 60:40 split respectively. The dimensions are in the table below.

**Table 1: Size of Various Data**

|  |  |  |  |
| --- | --- | --- | --- |
| **Unlabeled Data** | **Labeled Data** | **Train Data** | **Test Data** |
| 187849 | 272762 | 163657 | 109105 |

Next, I investigated which part of the data set had missing values. Table 2 below shows the missing values by each category. Some categories have missing data in over 20% of the data.

In this situation, I could either choose a model which deals effectively with missing values or delete those values for which there is data missing (i.e. Bayesian Networks). If I deleted all entries with at least one missing value, there would be 57,307 values left over or 35% of the original training data. This is probably enough to write a predictive value on. To be absolutely sure, I would have to analyze the VC dimension of the model. To be clear, removing missing values definitely introduces bias into the model. One way we might reduce this bias is to sample randomly from the remainder of the training data. However, a better strategy can be seen when we realize that data is only missing from 17 variables, so we can simply use the variables that are clean (tables 2 and 3).

**Table 2: Missing Values by Category in the Training Data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **agep** | **bld** | **cit** | **cow** | **dis** |
| 0 | 5806 | 0 | 67526 | 0 |
| **fes** | **fs** | **hht** | **hincp** | **indp** |
| 30889 | 0 | 5806 | 5806 | 67526 |
| **jwtr** | **mar** | **mil** | **msp** | **mv** |
| 92043 | 0 | 34794 | 30403 | 948 |
| **noc** | **np** | **pap** | **puma** | **rac1p** |
| 5806 | 0 | 30403 | 0 | 0 |
| **rwat** | **sch** | **semp** | **sex** | **st** |
| 5806 | 5700 | 30403 | 0 | 0 |
| **type** | **veh** | **wif** | **wkhp** | **hicov** |
| 0 | 5806 | 30889 | 80288 | 0 |

**Table 3: Missing Values by Category in the prediction data**

**The missing data is consistent**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **agep** | **bld** | **cit** | **cow** | **dis** |
| 0 | 6928 | 0 | 76351 | 0 |
| **fes** | **fs** | **hht** | **hicov** | **hincp** |
| 35712 | 0 | 6928 | 187849 | 6928 |
| **indp** | **jwtr** | **mar** | **mil** | **msp** |
| 76351 | 105750 | 0 | 38896 | 33794 |
| **mv** | **noc** | **np** | **pap** | **puma** |
| 1117 | 6928 | 0 | 33794 | 0 |
| **rac1p** | **rwat** | **sch** | **semp** | **sex** |
| 0 | 6928 | 6163 | 33794 | 0 |
| **st** | **type** | **veh** | **wif** | **wkhp** |
| 0 | 0 | 6928 | 35712 | 91685 |

My understanding is that variables with high correlation can be bad for logistic regressions. Usually, I would go through and remove variables that have low coefficients to try to reduce this problem. However, for a quick and dirty model, this should be reasonable.

In future iterations, I will clean up this model and also build a Bayesian net, which would be more appropriate for this dataset because since 70% of the data has at least one entry missing. I will also build ROC curves to more properly compare the models.