The data file concerns an insurance company that wants to build a model for flagging suspicious, potentially fraudulent claims. Variable Details:

|  |  |
| --- | --- |
| claim\_type | Type of claim (1 to 5) indicating First Party Insurance Claims  General Claims, Car Insurance Claims, Health Insurance Claims  Homeowner Claims |
| claim\_amount | Amount claimed |
| coverage | Amount covered |
| townsize | Town size of claimant |
| gender | Gender of claimant |
| edcat | Education of claimant ( 1 to 5) indicating uneducated, high school, undergraduate degree, graduate degree and professional. |
| retire | Retired or Not ( 1or 0) |
| income | Income of the claimant |
| marital | Marital Status of the claimant |
| reside | Residential status –categorical variable indicating status (e.g. non-resident/ Resident and Ordinarily Resident (ROR) etc) |
| primary\_residence | Indicates if the claimant filed from primary residence ( 1or 0) |
| fraud | 1 indicates fraudulent claim |

1. Partition of the data into Training and Test samples.
2. Fit a neural network model to classify “fraud.”
3. Examine the “classification confusion matrix” for the validation data. What are your comments on this?
4. How do you know that the model has not “memorised” the data ? i.e. overfitting issue.
5. What is the next course of action?