There is a real dataset, “*tax-audit.jmp*”, originating from the Swedish tax authority (Skatteverket). It concerns income tax returns for persons who during the year sold mutual funds managed in a foreign country. In order to preserve the privacy of both the taxpayers and the asset management companies, all personally identifiable information has been removed and the names of the companies and the funds have also been changed.

This data set consists of 3,119 income tax returns represented as a table with the following columns:

|  |  |
| --- | --- |
| No | Id-number |
| Assessment | Has the value **Correct** if the income tax return is correct, and **Incorrect**, otherwise. **This is the field that we want to be able to predict.** |
| Age | The age of the tax payer |
| Income from business | **'Yes'** if the tax payer has income from a private business, `**No'** otherwise. |
| Taxable fortune | In Sweden, there is a tax on personal net assets of more than 1.5 MSEK.  This column has the value `**Yes'** if the person has reported this. |
| Managed by Nowhere A.M. | **'Yes'** if the fund that the tax payer has sold was managed by the company  `Nowhere Asset Management'. This may be of interest since this company has given better information to the tax payer about the duty of reporting than other companies. |
| Sex | The sex of the tax payer. |
| Type of form | There are two types of income-tax return forms: `**Simplified'** is a simplified form where the tax authorities have filled in all numbers reported for the person by employers, banks, etc, and `**Complete'** is a form where nothing is filled in. |
| Income from employment | Income from employment |
| Fund company | The name of the company who managed the fund (the names have been changed). |
| Fund name | The name of the fund that was sold (the names have been changed). |
| Reported by | The name of the management firm who reported the sale to the tax authority (the names have been changed) |
| Prot/Loss | The amount of the realized profit or loss that was the result of the sale (in Swedish, Krona, SEK). Positive numbers are prot and negative numbers represent loss. |
| Validation | A column randomly generated with 80% rows of zero and 20% rows of one. This is used to partition the data. |

1. Create an initial logistic regression model to predict “Assessment” over **the whole data**. This initial model should include all independent variables EXCEPT “No”, “Fund company”, “Fund name” and “Reported by”. Manually apply backward stepwise selection to eliminate insignificant variables until all remaining variables have a P-value<0.05. (Note: Use the P-value displayed in the Effects Summary table.) Save the final model script to data file. Which variables are selected in the end?

2. According to your final model, answer the following questions. You only need to answer Yes or No. (Hint: Be careful about the alphabetical order of “Correct” and “Incorrect”.)

(a) Are older people more likely to file a correct tax return?

(b) Is a man more likely to file a correct tax return than a woman?

(c) Is the tax payer whose fund was managed `Nowhere Asset Management' more likely to file a correct tax return than others?

(d) Is a “Simplified” form more likely to be correct than the “Complete” one?

3. Use “Validation” column to partition the data into training set and validation set. Create a new final model as in step 1 with the target variable of **Incorrect** to answer the remaining questions. Display the confusion matrix.

What are the accuracy, precision and recall values of your model over validation set?

4. Generate the ROC curve. Using the validation data set: What is the area under the curve? Using the table from class, what is your assessment of it?

Save the script to data file.