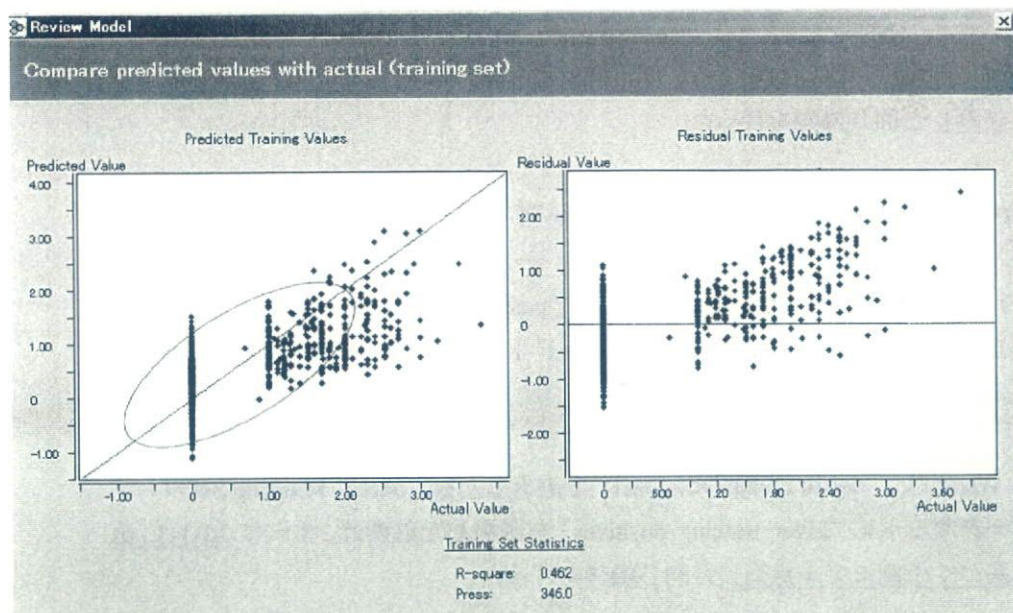


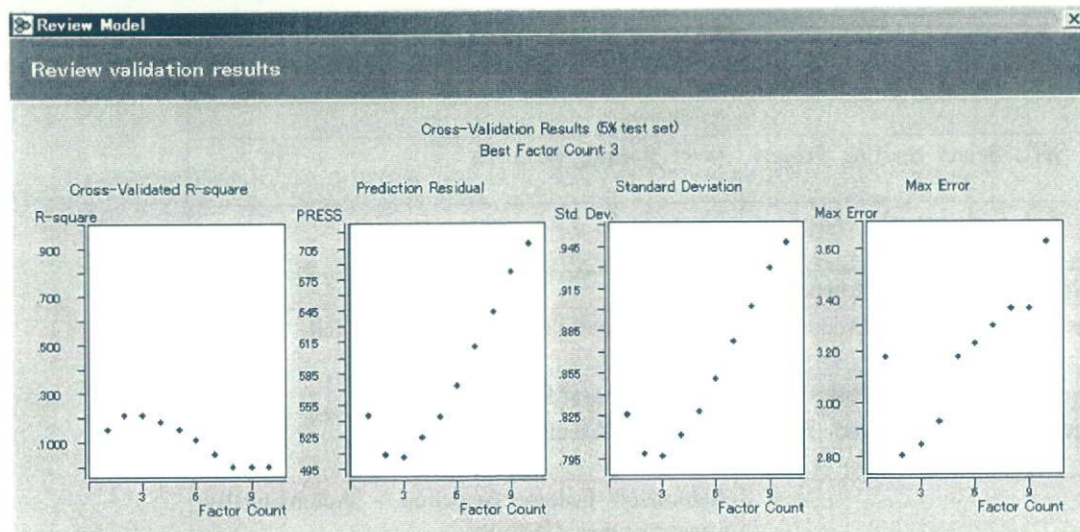
Test Set Statistics: Max Error

2.84

## 予測モデルと実データとの比較



## Cross-Validation Results



- (3) 構築したModelのR-squareの値が低いため、精度の高いModelとはいえない。Liver=0のデータが非常に多いことが原因と考えられる。そこで、Liver=0のデータを除いたデータセットでModel構築を試みた。

### 3-1-5. 肝毒性が陰性の化合物のみを除いたデータセットに基づく予測モデルの構築

トレーニングデータセットの読み込みについては、以下の変更点以外はすべて「3-1-2. トレーニングデータセットの読み込み」と同じ操作を行った。

- (1) Select Content : Liver=0のものを選択し”Delete Selected Rows”をクリックし削除した後、**Next**ボタンをクリックした。
- (2) Select a Database : Create a new Source ID:に “Liver\_training\_positives” を指定した。
- (3) Import Structures and Data : project名 “Liver\_training\_positives” にして、**Finish**ボタンをクリックして、データのインポートを行った。

手順は、Leadscope Task Wizardで、“Build Predictive Model” を選択した後、**Select Training Set**のステップでトレーニングデータセットに “Liver\_training\_positives” を選択した以外は、すべて「3-1-3. 全化合物データセットに基づく予測モデル構築」と同じ操作を行った。

### 3-1-6. 予測モデル構築結果

以下にモデル構築時の条件サマリーを記載する。

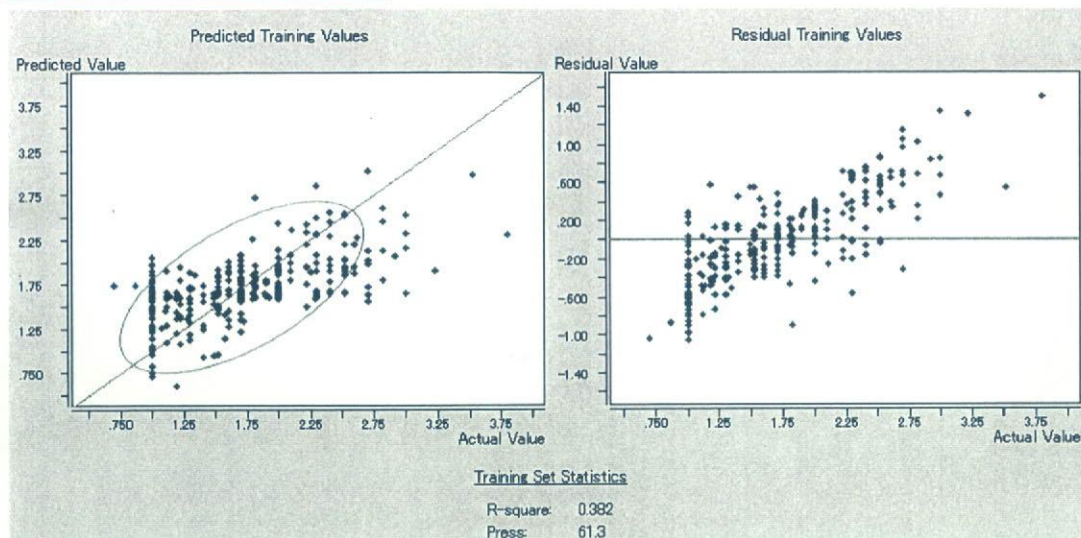
| 項目                                                                 | 選択条件                                                                  |
|--------------------------------------------------------------------|-----------------------------------------------------------------------|
| <b>Select Training Set</b> : Select Starting Project for Training  | Liver_training_positives                                              |
| <b>Select Training Set</b> : Select the response variables         | Liver (3)                                                             |
| <b>Select Training Set</b> : Edit Training Structures              | (Default)                                                             |
| <b>Build Knowledge</b> : Add or Create Additional Structural Rules | Rule名 : Scaffold Generation (Default)                                 |
| <b>Build Knowledge</b> : Filter Structural Rules                   | Filter Structural Rules (Default)                                     |
| <b>Build Knowledge</b> : Review Structures Rules Template          | No Edit (Default)                                                     |
| <b>Select Features</b> :                                           | Automated Feature Selection - Automatically select features (Default) |
| <b>Build Model</b> :                                               | Automated Model Building (Default)                                    |

- (4) 以下にModel構築の結果を記載した。

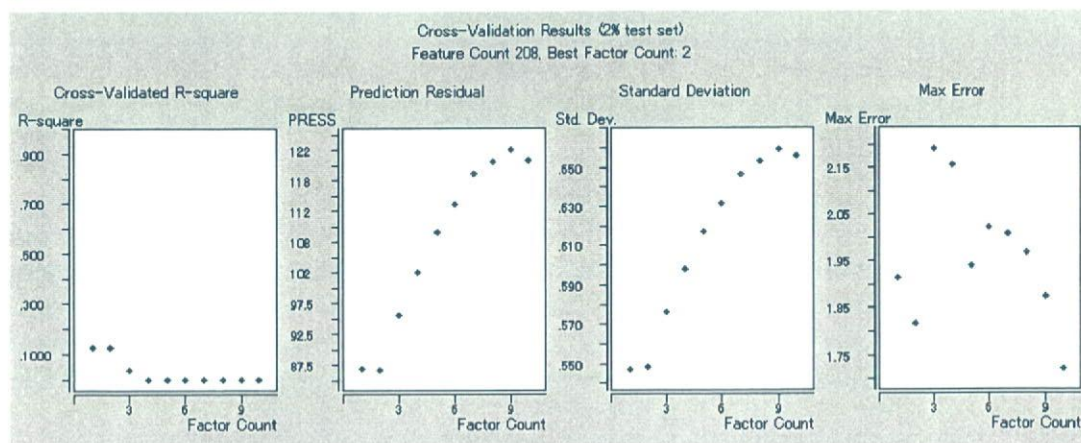
| 項目                    | 選択条件                             |
|-----------------------|----------------------------------|
| Build Knowledge       | 1 Additional Rule Set            |
| Select Features       | 208 Selected Features            |
| Prediction Model Name | PredictiveModel                  |
| Type                  | Partial Least Squares Regression |
| Factor Count          | 2                                |

|                                    |       |
|------------------------------------|-------|
| Training Set Statistics: R-squares | 0.382 |
| Training Set Statistics: Press     | 61.3  |
| Test Set Statistics: R-squares     | 0.125 |
| Test Set Statistics: Press         | 86.8  |
| Test Set Statistics: Std. Dev.     | 0.548 |
| Test Set Statistics: Max Error     | 1.82  |

### 予測モデルと実データとの比較



### Cross-Validation Results



- (5) 陰性データを除いたトレーニングデータセットからモデルを構築したが、Training Setに対するR-squareの値は、全データセットから構築したモデルの0.462から0.382に低下した。ただし、PRESS（残差の総和）は346.0から61.3へと大幅に減少したため、全体として予測精度は改善されている。また、上図「予測モデルと実データとの比較」の右図のResidual Training Valuesでは、

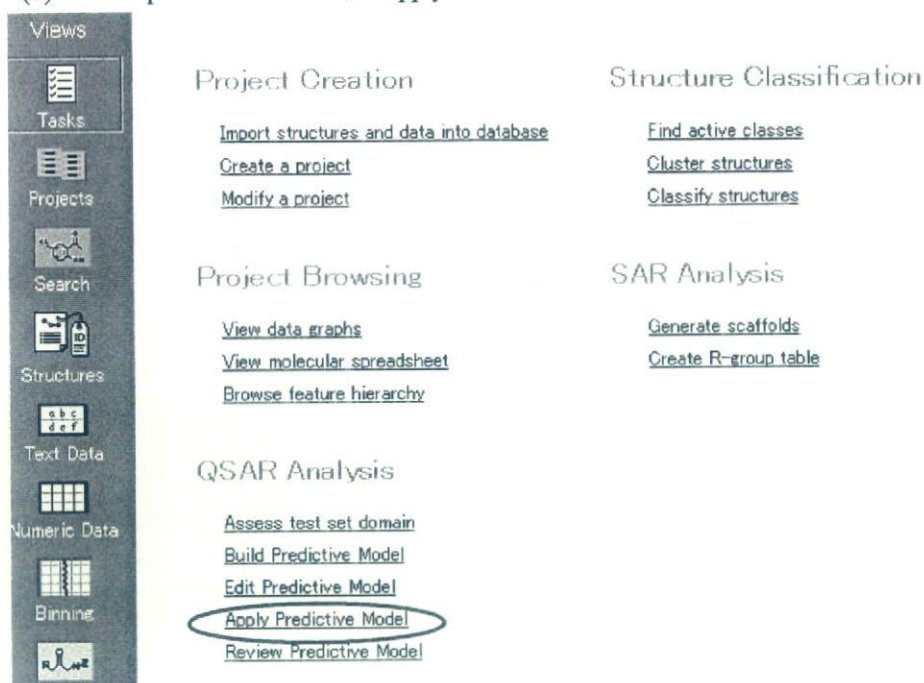
妥当なモデルであればResidual Value=0の近傍でランダムに散布図が作成されるべきだが実際にはそうっておらず偏りがある。

次にこのモデルの適用方法について”Apply Predictive Model”機能を使って検討した。

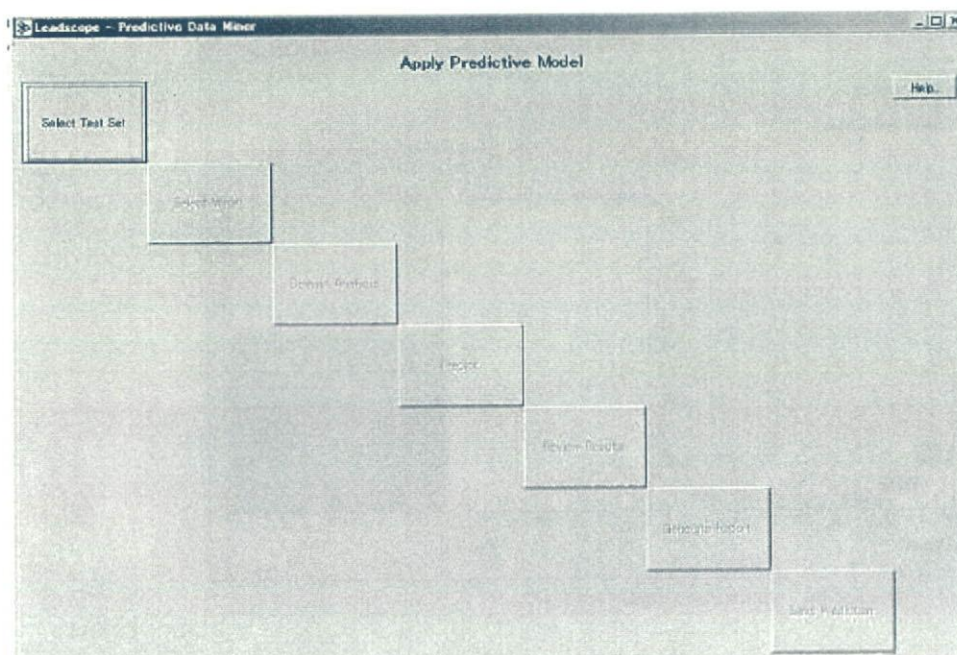
## 3-1-7. 予測モデル適用方法の検討

808全化合物に対して、3-1-6で構築したモデルを適用し予測精度を検討した。

(1) Leadscope Task Wizardで、“Apply Predictive Model”を選択した（下図）。



(2) “Select Test Set”をクリックした（下図）。

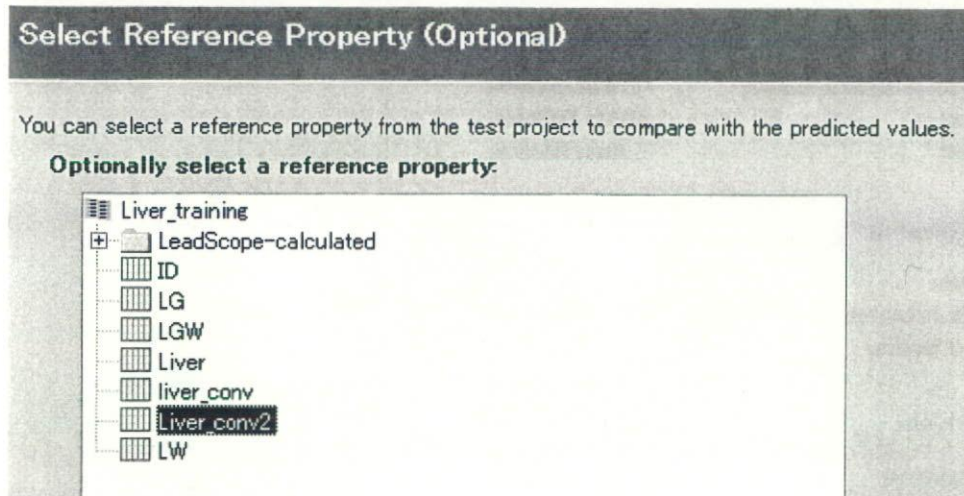


(3) Select Test Set Project: Projects -> Personal -> Liver\_trainingを選択。

(4) Select Model: PredictiveModelを選択。以下次項まではすべてデフォルト条件で進めた。

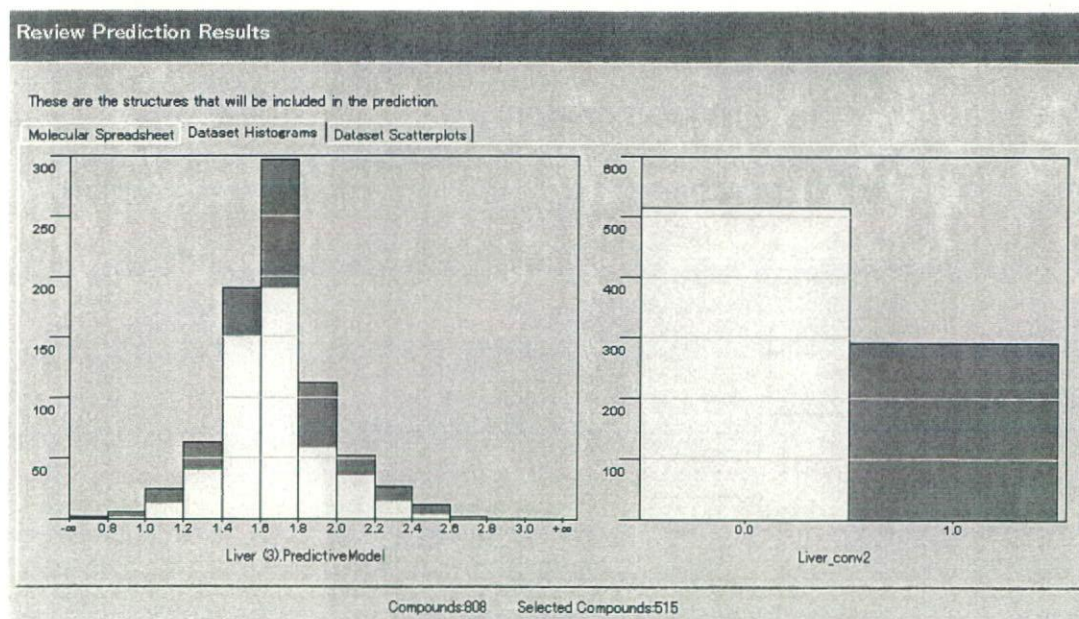
(5) Predict:

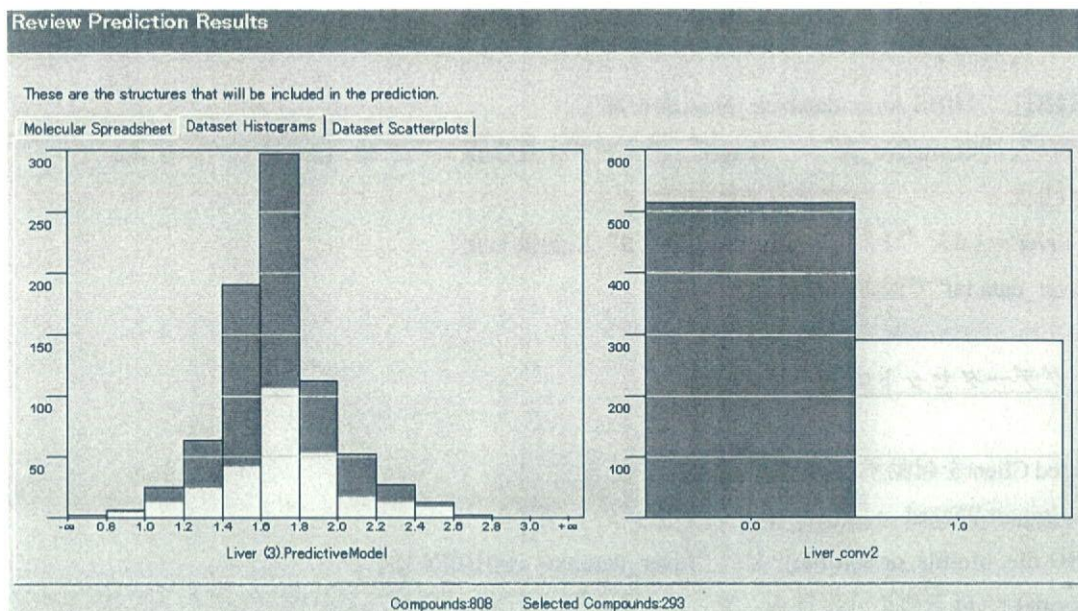
Select Reference Property: 予測結果と比較するための毒性データとして、Liver\_conv2(連続データをバイナリ化したデータ)を選択した(下図)。Liver\_conv2については「条件2: トレーニングデータセットの変更」を参照。



(6) Review Results:

下図の左のグラフが予測モデルによって予測された毒性値(Liver)のヒストグラムとなる。右のグラフは実際の毒性データ(Liver\_conv2)のヒストグラムとなる。双方の図において黄色の箇所が対応している。





#### (7) 考察

Liver\_conv2=0に対応する左側のヒストグラムとLiver\_conv2=1に対応する左側のヒストグラムとが分かれていれば、その境界値を使い陰性と陽性を判別することができるが、実際には陰性と陽性との分布は分かれなかったため、この予測モデルの活用は難しいと考えられる。

### 3-2. 毒性試験データにバイナリ値を用いた場合の予測モデル構築

- (1) ISIS/Baseを起動し、NIHS Ames database\_New.dbを開く。
- (2) 全件検索を行い、“Chem\_No\_A”・“Liver”フィールドをSARテーブル（Excelスプレッドシート）に出力した。
- (3) Excel上で、Liver $\geq$ 1.6を“1”に、Liver $<$ 1.6を“0”に変換した。
- (4) 変換後、“Liver\_data.txt”で保存した。

#### 3-2-1. トレーニングデータセットの読み込み

- (1) Leadscope Hosted Clientを起動する。
- (2) Welcome to the Import Wizard : **Next** ボタンクリック
- (3) Import a file (SD file, Molfile, or text file): に “Liver\_data.txt” を指定する。
- (4) Select a File Format : 以下のように選択した。

Select a File Format

Select file format: **Delimited Text**

Data File Format Options

Select the delimiter character separating the text fields:

☒ tab      ☐ comma      ☐ other

☐ space      ☐ semicolon

Other:

☒ The first row contains data set names

Structure ID Format Options

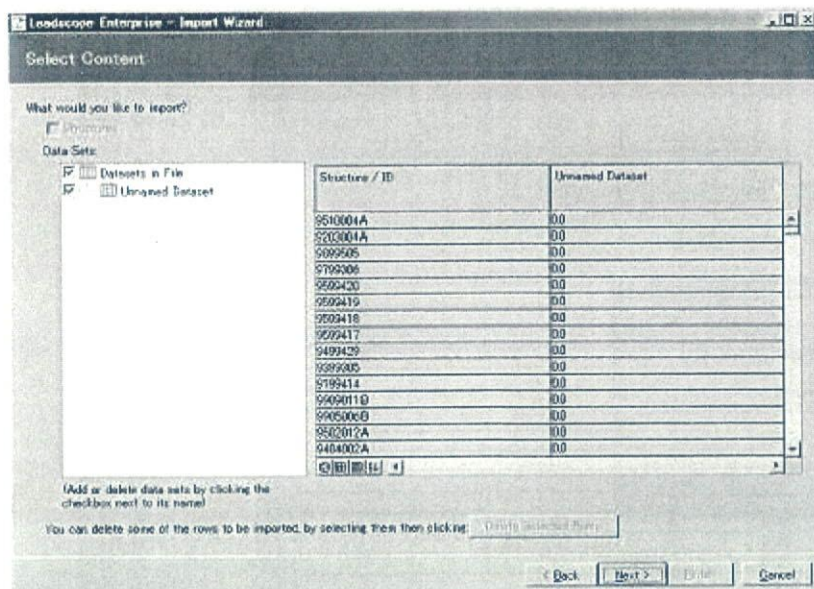
The structure ID is:

☐ on the first line      ☐ in a field with name:

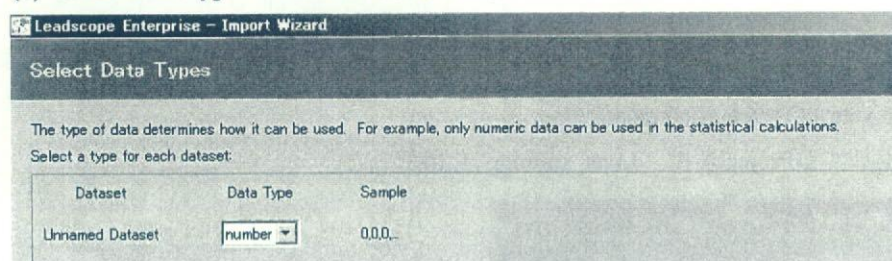
File Preview (this only contains the first few entries in the file):

| Structure / ID | Unnamed Dataset |
|----------------|-----------------|
| 8811007A       | 0.0             |
| 8812012A       | 0.0             |
| 8903001A       | 0.0             |
| 8903002A       | 0.0             |
| 8903008A       | 0.0             |
| 8903010A       | 0.0             |
| 8903012A       | 0.0             |
| 8903014A       | 0.0             |
| 8905007B       | 0.0             |
| 8905010A       | 0.0             |

- (5) Select Content : インポートするデータ内容を確認した。

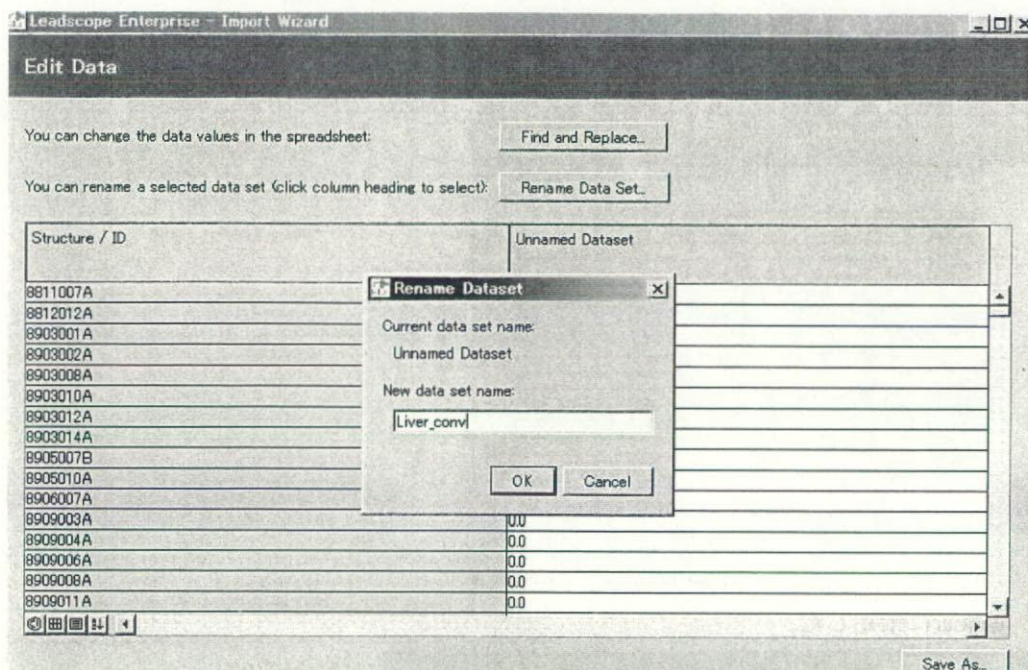


(6) Select Data Types : numberに指定した。

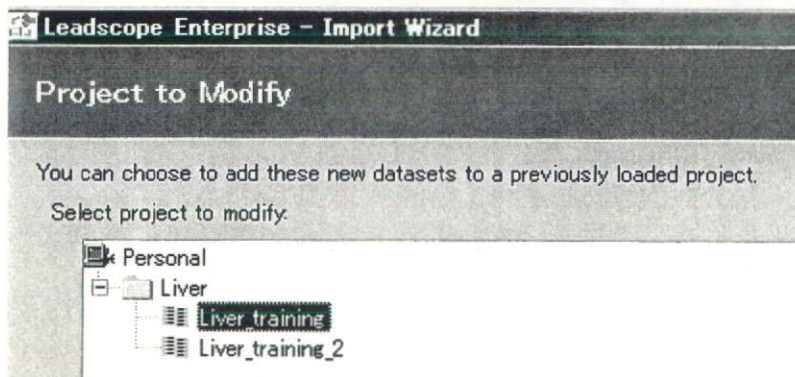


(7) Edit Structure IDs : 編集は行わずNextボタンをクリックした。

(8) Edit Data : Unnamed Datasetを“Liver\_conv”にRenameした。



- (9) Calculate New Numeric Datasets : 編集は行わずNextボタンをクリックした。
- (10) Select a Database : database on your PC (Default)を選択した。
- (11) Project to Modify : データを追加するProject名は “Liver\_training”を選択した。



- (12) Binning Schemes For New Datasets : 下記のように設定した。

**Leadscope Enterprise - Import Wizard**

**Binning Schemes For New Datasets**

For each new data set, specify whether a new or existing binning scheme should be assigned.

|                                                          |                                                                                 |
|----------------------------------------------------------|---------------------------------------------------------------------------------|
| Liver_conv                                               | Binning Scheme: Autobinned-Liver_conv                                           |
| <input checked="" type="radio"/> Generate binning scheme | <input checked="" type="radio"/> Higher values indicate a stronger correlation. |
| <input type="radio"/> Use existing binning scheme        | <input type="radio"/> Lower values indicate a stronger correlation.             |

(13) Import Structures And Data : 確認して **Finish** ボタンをクリックした。

### 3-2-2. 予測モデルの構築

条件1 : Default条件でのモデル構築結果 (Liver conv) :

モデル構築の際の各条件は以下の表の通りに選択した。

| 項目                                                                 | 選択条件                                                                  |
|--------------------------------------------------------------------|-----------------------------------------------------------------------|
| <b>Select Training Set</b> : Select Starting Project for Training  | Liver_training                                                        |
| <b>Select Training Set</b> : Select the response variables         | Liver_conv                                                            |
| <b>Select Training Set</b> : Edit Training Structures              | (Default)                                                             |
| <b>Build Knowledge</b> : Add or Create Additional Structural Rules | Rule名 : Scaffold Generation(6) (Default)                              |
| <b>Build Knowledge</b> : Filter Structural Rules                   | Filter Structural Rules (Default)                                     |
| <b>Build Knowledge</b> : Review Structures Rules Template          | No Edit (Default)                                                     |
| <b>Select Features</b> :                                           | Automated Feature Selection - Automatically select features (Default) |
| <b>Build Model</b> :                                               | Automated Model Building (Default)                                    |

<Model構築の結果>

| 項目                                   | 選択条件                              |
|--------------------------------------|-----------------------------------|
| Build Knowledge                      | 95 Rules (Scaffold Generation(6)) |
| Select Features                      | 207 Selected                      |
| Prediction Model Name                | Predictive Model (Run#10)         |
| Type                                 | Partial Logistic Regression       |
| Factor Count                         | 3                                 |
| Training Set Statistics: Concordance | 87.6                              |
| Training Set Statistics: Sensitivity | 52.1                              |
| Training Set Statistics: Specificity | 96.9                              |

|                                        |      |
|----------------------------------------|------|
| Training Set Statistics: ROC Positives | 4.3  |
| Training Set Statistics: ROC Negatives | 7.8  |
| Test Set Statistics: Concordance       | 83.8 |
| Test Set Statistics: Sensitivity       | 42.4 |
| Test Set Statistics: Specificity       | 94.5 |
| Test Set Statistics: ROC Positives     | 2.0  |
| Test Set Statistics: ROC Negatives     | 6.3  |

条件2 : Featuresを選択した条件でのモデル構築結果 (Liver conv) :

<条件2のモデル構築条件>

上記Defaultのモデルに対してEdit Predictive Modelを利用してFeaturesを選択して作成した。

| 項目                                                          | 選択条件                                     |
|-------------------------------------------------------------|------------------------------------------|
| Select Training Set : Select Starting Project for Training  | Liver_training                           |
| Select Training Set : Select the response variables         | Liver_conv                               |
| Select Training Set : Edit Training Structures              | (Default)                                |
| Build Knowledge : Add or Create Additional Structural Rules | Rule名 : Scaffold Generation(6) (Default) |
| Build Knowledge : Filter Structural Rules                   | Filter Structural Rules (Default)        |
| Build Knowledge : Review Structures Rules Template          | No Edit (Default)                        |
| Select Features :                                           | Manually Selected Features<br>下記参照       |
| Build Model :                                               | Automated Model Building (Default)       |

Mean Liver\_convが0.3以上～0.6未満のFeatureを削除した。

# Appendix 9

These are the features that will be included in the model.

| Name                             | Frequency | liver_conv | Z-Score:<br>liver_conv | Chi-Square:<br>liver_conv | Mean:<br>liver_conv | Std Dev:<br>liver_conv | liver_conv |
|----------------------------------|-----------|------------|------------------------|---------------------------|---------------------|------------------------|------------|
| benzene, 1,3-dimethyl-           | 18        |            | 1.947                  | 2.733                     | 0.3889              | 0.5016                 |            |
| benzene, 1-ethyl-                | 10        |            | 1.531                  | 1.296                     | 0.4                 | 0.5164                 |            |
| methane, 1-aryl-,1-hydroxy-      | 10        |            | 1.531                  | 1.296                     | 0.4                 | 0.5164                 |            |
| sulfone, aryl-                   | 10        |            | 1.531                  | 1.296                     | 0.4                 | 0.5164                 |            |
| nitrile, aryl-                   | 20        |            | 2.179                  | 3.612                     | 0.4                 | 0.5026                 |            |
| Scaffold 153                     | 10        |            | 1.531                  | 1.296                     | 0.4                 | 0.5164                 |            |
| benzene, 1-alkyl-,4-halo-        | 17        |            | 2.127                  | 3.329                     | 0.4118              | 0.5073                 |            |
| sulfone                          | 12        |            | 1.823                  | 2.145                     | 0.4167              | 0.5149                 |            |
| nitrile, alkyl, acyc-            | 12        |            | 1.823                  | 2.145                     | 0.4167              | 0.5149                 |            |
| Scaffold 265                     | 12        |            | 1.823                  | 2.145                     | 0.4167              | 0.5149                 |            |
| Scaffold 131                     | 12        |            | 1.823                  | 2.145                     | 0.4167              | 0.5149                 |            |
| Scaffold 213                     | 16        |            | 2.319                  | 4.032                     | 0.4375              | 0.5123                 |            |
| benzene, 1-fluoro-,4-halo-       | 9         |            | 1.783                  | 1.876                     | 0.4444              | 0.527                  |            |
| Scaffold 84                      | 9         |            | 1.783                  | 1.876                     | 0.4444              | 0.527                  |            |
| Scaffold 52                      | 9         |            | 1.783                  | 1.876                     | 0.4444              | 0.527                  |            |
| Scaffold 111                     | 20        |            | 2.739                  | 6.056                     | 0.45                | 0.5104                 |            |
| Scaffold 167                     | 11        |            | 2.058                  | 2.833                     | 0.4545              | 0.5222                 |            |
| Scaffold 70                      | 11        |            | 2.058                  | 2.833                     | 0.4545              | 0.5222                 |            |
| benzene, 1,2-dihalo-             | 26        |            | 3.283                  | 9.233                     | 0.4615              | 0.5084                 |            |
| benzene, 1-cyclohexyl-           | 13        |            | 2.303                  | 3.834                     | 0.4615              | 0.5189                 |            |
| Scaffold 121                     | 26        |            | 3.283                  | 9.233                     | 0.4615              | 0.5084                 |            |
| Scaffold 98                      | 13        |            | 2.303                  | 3.834                     | 0.4615              | 0.5189                 |            |
| 1-benzene-nitrile                | 15        |            | 2.526                  | 4.862                     | 0.4667              | 0.5164                 |            |
| Scaffold 110                     | 15        |            | 2.526                  | 4.862                     | 0.4667              | 0.5164                 |            |
| benzene, 1-alkoxy-,4-alkyl-      | 12        |            | 2.543                  | 4.772                     | 0.5                 | 0.5222                 |            |
| benzene, 1-alkoxy-,4-amino-      | 6         |            | 1.791                  | 1.652                     | 0.5                 | 0.5477                 |            |
| benzene, 1-alkyl-,3-halo-        | 16        |            | 2.944                  | 6.933                     | 0.5                 | 0.5164                 |            |
| methane, 1-hydroxy-,1-phenyl-    | 8         |            | 2.071                  | 2.665                     | 0.5                 | 0.5345                 |            |
| benzene, 1-hydroxymethyl-        | 10        |            | 2.318                  | 3.71                      | 0.5                 | 0.527                  |            |
| Scaffold 221                     | 6         |            | 1.791                  | 1.652                     | 0.5                 | 0.5477                 |            |
| Scaffold 216                     | 6         |            | 1.791                  | 1.652                     | 0.5                 | 0.5477                 |            |
| Scaffold 136                     | 10        |            | 2.318                  | 3.71                      | 0.5                 | 0.527                  |            |
| Scaffold 74                      | 6         |            | 1.791                  | 1.652                     | 0.5                 | 0.5477                 |            |
| Scaffold 50                      | 6         |            | 1.791                  | 1.652                     | 0.5                 | 0.5477                 |            |
| thiophene                        | 9         |            | 2.613                  | 4.837                     | 0.5556              | 0.527                  |            |
| Scaffold 113                     | 9         |            | 2.613                  | 4.837                     | 0.5556              | 0.527                  |            |
| Scaffold 17                      | 9         |            | 2.613                  | 4.837                     | 0.5556              | 0.527                  |            |
| Scaffold 272                     | 7         |            | 2.406                  | 3.753                     | 0.5714              | 0.5345                 |            |
| oxycarbonyl, O-(t-butyl)-        | 5         |            | 2.189                  | 2.674                     | 0.6                 | 0.5477                 |            |
| benzene, 1-amino(NH2)-,3-methyl- | 5         |            | 2.189                  | 2.674                     | 0.6                 | 0.5477                 |            |

Features:207   Selected Features:51   Compounds:808   Selected Compounds:323

## <Model構築の結果>

| 項目                                     | 選択条件                              |
|----------------------------------------|-----------------------------------|
| Build Knowledge                        | 69 Rules (Scaffold Generation(6)) |
| Select Features                        | 156 Selected                      |
| Prediction Model Name                  | Predictive Model (Run#11)         |
| Type                                   | Partial Logistic Regression       |
| Factor Count                           | 6                                 |
| Training Set Statistics: Concordance   | 87.3                              |
| Training Set Statistics: Sensitivity   | 43.6                              |
| Training Set Statistics: Specificity   | 98.6                              |
| Training Set Statistics: ROC Positives | 8.0                               |
| Training Set Statistics: ROC Negatives | 6.74                              |
| Test Set Statistics: Concordance       | 85.3                              |
| Test Set Statistics: Sensitivity       | 40.0                              |
| Test Set Statistics: Specificity       | 97.0                              |
| Test Set Statistics: ROC Positives     | 3.47                              |
| Test Set Statistics: ROC Negatives     | 6.23                              |

条件3 : Default条件でのモデル構築結果 (Liver\_conv2) :

Liver =0を“0”に、それ以外は“1”に変換後、“Liver\_data2.txt”で保存した。その値をLeadscopeに読み込み、Liver\_conv2データとし、その値を用いて予測モデル構築を行った。

< Liver\_conv2を用いたDefault条件のモデル構築条件>

上記Defaultのモデルに対してBuild Predictive Modelを利用してFeaturesを選択して作成した。

| 項目                                                          | 選択条件                                                                  |
|-------------------------------------------------------------|-----------------------------------------------------------------------|
| Select Training Set : Select Starting Project for Training  | Liver_training                                                        |
| Select Training Set : Select the response variables         | Liver_conv2                                                           |
| Select Training Set : Edit Training Structures              | (Default)                                                             |
| Build Knowledge : Add or Create Additional Structural Rules | Rule名 : Scaffold Generation(7) (Default)                              |
| Build Knowledge : Filter Structural Rules                   | Filter Structural Rules (Default)                                     |
| Build Knowledge : Review Structures Rules Template          | No Edit (Default)                                                     |
| Select Features :                                           | Automated Feature Selection - Automatically select features (Default) |
| Build Model :                                               | Automated Model Building (Default)                                    |

<Model構築の結果>

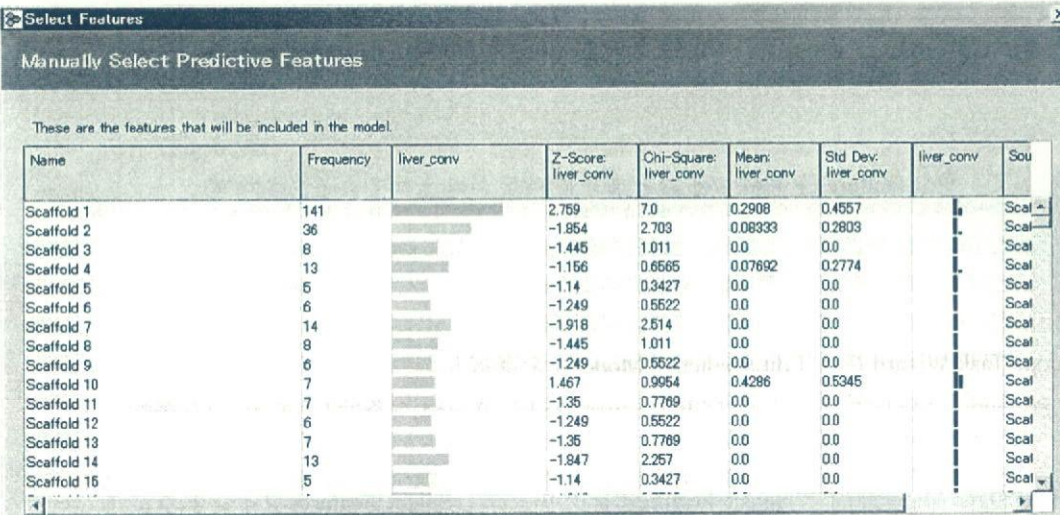
| 項目                                     | 選択条件                               |
|----------------------------------------|------------------------------------|
| Build Knowledge                        | 106 Rules (Scaffold Generation(7)) |
| Select Features                        | 207 Selected                       |
| Prediction Model Name                  | Predictive Model (Run#12)          |
| Type                                   | Partial Logistic Regression        |
| Factor Count                           | 3                                  |
| Training Set Statistics: Concordance   | 77.5                               |
| Training Set Statistics: Sensitivity   | 53.4                               |
| Training Set Statistics: Specificity   | 91.4                               |
| Training Set Statistics: ROC Positives | 3.55                               |
| Training Set Statistics: ROC Negatives | 3.42                               |
| Test Set Statistics: Concordance       | 74.4                               |
| Test Set Statistics: Sensitivity       | 49.0                               |
| Test Set Statistics: Specificity       | 89.0                               |
| Test Set Statistics: ROC Positives     | 2.55                               |
| Test Set Statistics: ROC Negatives     | 3.04                               |

条件4 : FeaturesをManualで取捨選択した条件でのモデル構築結果 (Liver\_conv2) :

| 項目                                                          | 選択条件                                     |
|-------------------------------------------------------------|------------------------------------------|
| Select Training Set : Select Starting Project for Training  | Liver_training                           |
| Select Training Set : Select the response variables         | Liver_conv2                              |
| Select Training Set : Edit Training Structures              | (Default)                                |
| Build Knowledge : Add or Create Additional Structural Rules | Rule名 : Scaffold Generation(4) (Default) |
| Build Knowledge : Filter Structural Rules                   | Filter Structural Rules (Default)        |
| Build Knowledge : Review Structures Rules Template          | No Edit (Default)                        |
| Select Features :                                           | Manual Selected Features<br>※ 参照         |
| Build Model :                                               | Automated Model Building (Default)       |

※参照

下記画面でFeaturesの選択を行った。



Manually Select Predictive Features

These are the features that will be included in the model.

| Name        | Frequency | liver_conv | Z-Score: liver_conv | Chi-Square: liver_conv | Mean: liver_conv | Std Dev: liver_conv | liver_conv | Sou  |
|-------------|-----------|------------|---------------------|------------------------|------------------|---------------------|------------|------|
| Scaffold 1  | 141       |            | 2.759               | 7.0                    | 0.2908           | 0.4557              |            | Scal |
| Scaffold 2  | 36        |            | -1.854              | 2.703                  | 0.08933          | 0.2803              |            | Scal |
| Scaffold 3  | 8         |            | -1.445              | 1.011                  | 0.0              | 0.0                 |            | Scal |
| Scaffold 4  | 13        |            | -1.156              | 0.6565                 | 0.07692          | 0.2774              |            | Scal |
| Scaffold 5  | 5         |            | -1.14               | 0.3427                 | 0.0              | 0.0                 |            | Scal |
| Scaffold 6  | 6         |            | -1.249              | 0.5522                 | 0.0              | 0.0                 |            | Scal |
| Scaffold 7  | 14        |            | -1.918              | 2.514                  | 0.0              | 0.0                 |            | Scal |
| Scaffold 8  | 8         |            | -1.445              | 1.011                  | 0.0              | 0.0                 |            | Scal |
| Scaffold 9  | 6         |            | -1.249              | 0.5522                 | 0.0              | 0.0                 |            | Scal |
| Scaffold 10 | 7         |            | 1.467               | 0.9954                 | 0.4286           | 0.5345              |            | Scal |
| Scaffold 11 | 7         |            | -1.35               | 0.7769                 | 0.0              | 0.0                 |            | Scal |
| Scaffold 12 | 6         |            | -1.249              | 0.5522                 | 0.0              | 0.0                 |            | Scal |
| Scaffold 13 | 7         |            | -1.35               | 0.7769                 | 0.0              | 0.0                 |            | Scal |
| Scaffold 14 | 13        |            | -1.947              | 2.257                  | 0.0              | 0.0                 |            | Scal |
| Scaffold 15 | 5         |            | -1.14               | 0.3427                 | 0.0              | 0.0                 |            | Scal |

Features:309   Selected Features:0   Compounds:808   Selected Compounds:0

- Add Calculated Descriptorsを選択して、Leadscope Calculated Properties (AlogP・HBA・HBD・Lipinski Score・MW・Parent Atom Count・Parent Molecular Weight・PSA・RotBond)を選択。
- PositiveとNegative化合物が等量ずつ含まれるFeatureは重要でないと考えられ、**Mean liver\_conv**が、0.3~0.5714のFeaturesを“Remove Selected Features”で除外した。

## &lt;Model構築の結果&gt;

| 項目                                     | 選択条件                               |
|----------------------------------------|------------------------------------|
| Build Knowledge                        | 324 Rules (Scaffold Generation(4)) |
| Select Features                        | 624 Selected                       |
| Prediction Model Name                  | Predictive Model (Run#4)           |
| Type                                   | Partial Logistic Regression        |
| Factor Count                           | 2                                  |
| Training Set Statistics: Concordance   | 80.0                               |
| Training Set Statistics: Sensitivity   | 62.3                               |
| Training Set Statistics: Specificity   | 90.2                               |
| Training Set Statistics: ROC Positives | 3.64                               |
| Training Set Statistics: ROC Negatives | 4.17                               |
| Test Set Statistics: Concordance       | 75.5                               |
| Test Set Statistics: Sensitivity       | 55.5                               |
| Test Set Statistics: Specificity       | 87.0                               |
| Test Set Statistics: ROC Positives     | 2.45                               |
| Test Set Statistics: ROC Negatives     | 3.41                               |

条件5：条件4に対してMean liver\_convが0.1～0.6となるFeaturesを削除(Liver\_conv2)

## &lt;目的&gt;

条件4のモデルに対して、更にFeaturesを除外することで予測精度が向上するかを検討する。

## &lt;方法&gt;

## 予測モデルの構築

- (1) Leadscope Task Wizardで、“Edit Predictive Model”を選択した。
- (2) Select Predictive Modelウィンドウで”Predictive Model (Run#4)”(条件2のモデル)を選択した。
- (3) Predictive Data Miner画面で”Select Features”をクリックしSelect Featuresウィンドウにて Manually select featuresを選択した後、Nextボタンをクリックした。
- (4) Leadscope Calculated Properties (AlogP・HBA・HBD・Lipinski Score・Parent Atom Count・Parent Molecular Weight・PSA・RotBond)は既にモデルのFeaturesに含まれていたため、Add Calculated Descriptorsは使用しなかった。
- (5) PositiveとNegative化合物が等量ずつ含まれるFeatureは重要でないと考えられ、Mean liver\_convが、0.1以上～0.6未満のFeaturesを “Remove Selected Features” で除外した。
- (6) 以降は、今までと全く同様の手順で行った。

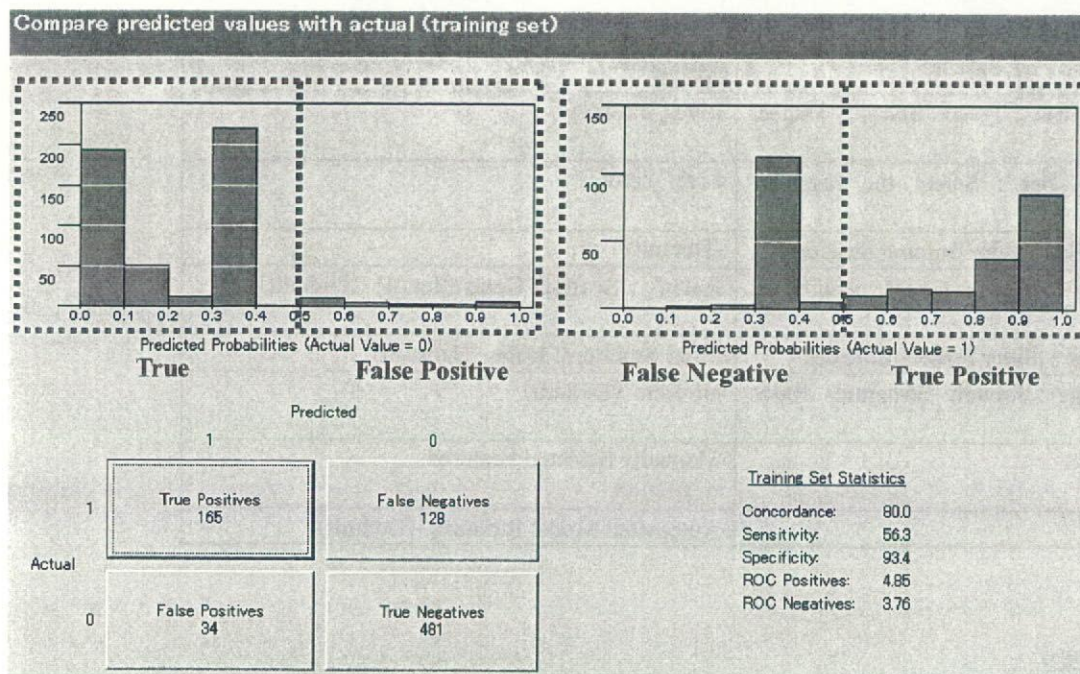
## &lt;条件5のモデル構築条件&gt;

| 項目                                                          | 選択条件                                     |
|-------------------------------------------------------------|------------------------------------------|
| Select Training Set : Select Starting Project for Training  | Liver_training                           |
| Select Training Set : Select the response variables         | Liver_conv2                              |
| Select Training Set : Edit Training Structures              | (Default)                                |
| Build Knowledge : Add or Create Additional Structural Rules | Rule名 : Scaffold Generation(4) (Default) |
| Build Knowledge : Filter Structural Rules                   | Filter Structural Rules (Default)        |
| Build Knowledge : Review Structures Rules Template          | No Edit (Default)                        |
| Select Features :                                           | Manually Selected Features<br>上記参照       |
| Build Model :                                               | Automated Model Building (Default)       |

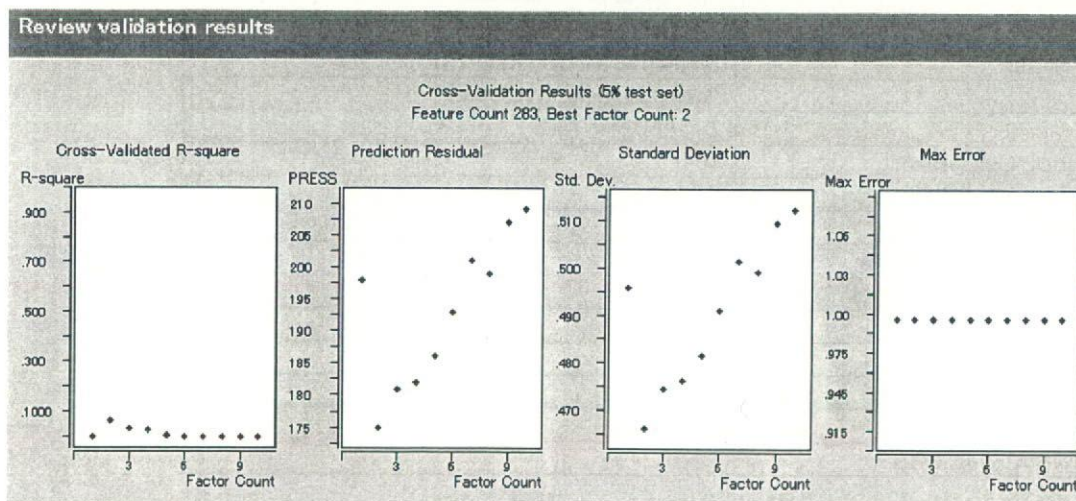
## &lt;Model構築の結果&gt;

| 項目                                     | 選択条件                               |
|----------------------------------------|------------------------------------|
| Build Knowledge                        | 123 Rules (Scaffold Generation(4)) |
| Select Features                        | 283 Selected                       |
| Prediction Model Name                  | Predictive Model (Run#13)          |
| Type                                   | Partial Logistic Regression        |
| Factor Count                           | 2                                  |
| Training Set Statistics: Concordance   | 80.0                               |
| Training Set Statistics: Sensitivity   | 56.3                               |
| Training Set Statistics: Specificity   | 93.4                               |
| Training Set Statistics: ROC Positives | 4.85                               |
| Training Set Statistics: ROC Negatives | 3.76                               |
| Test Set Statistics: Concordance       | 78.2                               |
| Test Set Statistics: Sensitivity       | 54.6                               |
| Test Set Statistics: Specificity       | 91.7                               |
| Test Set Statistics: ROC Positives     | 3.72                               |
| Test Set Statistics: ROC Negatives     | 3.55                               |

予測モデルの精度



### Cross-Validation Results



Mean Liver\_conv2を0.3~0.6から0.1~0.6に範囲を広げ、この範囲のFeaturesを除外してモデルを構築した場合、感度が下がってしまうことが分かった。この原因としては、Featuresが割り当てられなかったためにPredicted Probabilitiesが0.5以下の毒性化合物が増えたためfalse negative化合物が増えたためと考えられる（プログラムはPredicted Probabilitiesが0.5より小さい時は「陰性」で0.5より大きい時は「陽性」と判断している）。

Predicted Probabilitiesが0.3~0.4の範囲に予測された化合物はほとんどFeaturesのない化合物である。

これら以外の化合物に関しては非常に予測精度が高い。このため、今後はPredicted Probabilities $\leq 0.2$ →「陰性」、Predicted Probabilities $\geq 0.6$ →「陽性」、それ以外のケース→予測不能、という判別を下すモデルが作成できれば良いと考えられる。

以降では、予測モデルの感度を上げるためにGOT/GPT, Weight\_Liverのデータも考慮に入れてモデルを構築する。

### 3-2-3. 予測モデルのまとめ 1

| 条件  | Response variables | Scaffold Generation# | モデル名                      | Features | Feature selection                  | concordance | sensitivity | specificity |
|-----|--------------------|----------------------|---------------------------|----------|------------------------------------|-------------|-------------|-------------|
| 条件1 | Liver_conv         | 1 (95Rules)          | Predictive Model (Run#10) | 207      | Automatic                          | 87.1        | 50.3        | 96.7        |
| 条件2 | Liver_conv         | 6 (69Rules)          | Predictive Model (Run#11) | 156      | Manual(Meanが0.3-0.6となるFeaturesを除外) | 87.3        | 43.6        | 98.6        |
| 条件3 | Liver_conv2        | 7 (106Rules)         | Predictive Model (Run#12) | 207      | Automatic                          | 77.5        | 53.4        | 91.4        |
| 条件4 | Liver_conv2        | 4 (324Rules)         | Predictive Model (Run#4)  | 624      | Manual(Meanが0.3-0.6となるFeaturesを除外) | 80.0        | 62.3        | 90.2        |
| 条件5 | Liver_conv2        | 4 (123Rules)         | Predictive Model (Run#13) | 283      | Manual(Meanが0.1-0.6となるFeaturesを除外) | 80.0        | 56.3        | 93.4        |

### 3-4. GOT/GPT、肝重量のバイナリデータも考慮に入れたモデルの構築

#### 3-4-1. Liver、GOT/GPT、肝重量のバイナリデータファイルの作成

- (1) ISIS/Baseを起動し、NIHS Ames database\_New.dbを開く。
- (2) 全件検索を行い、{"Chem\_No\_A"・"Liver"・"GOT/GPT"・"Liver\_Weight" フィールド}をSARテーブル(Excelスプレッドシート)に出力した。
- (3) Excel上で、以下3通りでデータを変換した：
  - ①Liver、GOT/GPT、Liver\_Weightのいずれも0の場合“0”に、それ以外は“1”に変換した値をLGWとした。
  - ②Liver、GOT/GPTのいずれも0の場合“0”に、それ以外は“1”に変換した値をLGとした。
  - ③Liver、Liver\_Weightのいずれも0の場合“0”に、それ以外は“1”に変換した値をLWとした。
- (4) 変換後、各々を“LGW\_data.txt”、“LG\_data.txt”、“LW\_data.txt”として保存した。

#### 3-4-2. トレーニングデータセットの読み込み

手順は基本的には「3-5. トレーニングデータセットの読み込み」に従った。

- (1) Leadscope Hosted Clientを起動する。
- (2) Welcome to the Import Wizard :  ボタンをクリック
- (3) Import a file (SD file, Molfile, or text file): に “LGW\_data.txt” を指定する。
- (4) Select a File Format : デフォルトで設定した。
- (5) Select Content : インポートするデータ内容を確認した。
- (6) Select Data Types : numberに指定した。デフォルト。
- (7) Edit Structure IDs : 編集は行わず ボタンをクリックした。
- (8) Edit Data : Dataset名が予め“LGW”だったためデフォルトでNEXTボタンをクリックした。
- (9) Calculate New Numeric Datasets : 編集は行わず ボタンをクリックした。
- (10) Select a Database : database on your PC (Default)を選択した。
- (11) Specify Target Folders (optional) : 編集は行わず ボタンをクリックした
- (12) Project to Modify : データを追加するProject名として “Liver\_training”を選択した。
- (13) Binning Schemes For New Datasets : 下記のように設定した。