CHAPTER 12: EXERCISE 3

CLASSIFICATION TREES

NITIN KALÉ, UNIVERSITY OF SOUTHERN CALIFORNIA
NANCY JONES, SAN DIEGO STATE UNIVERSITY

OBJECTIVE

The objective of this exercise is to classify cases of equipment failure and use that to predict future failures.

ACTIVITIES

- Import and prepare data
- Apply data mining algorithms
- Configure forecasting models
- Create data visualizations
- Analyze and interpret output from models
- Publish results

SOFTWARE PREREQUISITES

- SAP Predictive Analytics 2.2
- Microsoft Excel

UCC PRODUCTS REQUIRED

- None

DATA SET

- Data file *transformer_failures_E12_3.xlsx*

SCENARIO

Equipment failure is an inevitable reality of all utility companies. Preventative maintenance is the best way to avoid costly failures, delays in customer service, and expensive repairs. Such maintenance can be done on a regular basis for all equipment but that may not be economical. 'Predictive' maintenance identifies which equipment is likely to fail in a given set of constraints (such as time period, event, and load). We would like to predict the failure probability based on historical data. We will use a decision tree to create a predictive model for preventative maintenance.

CLASSIFICATION TREES

Within predictive data mining methods, classification is a powerful way to classifying a target variable into a category; for instance to predict if equipment will fail based on attributes that most influence this outcome. The results of the data mining training exercise generate a decision tree. A decision tree is a hierarchical tree-shaped model that is used to determine how one choice at a node leads to the next with a statistical probability. The branches at each node are mutually exclusive.

Decision trees are particularly useful when the outcome of a decision cannot be predicted with certainty; also when the number of influencing factors is large.

You will use SAP Predictive Analytics – Automated Analytics to train the classification model.

To build an analysis for equipment failure, proceed as follows:

1. Open the Microsoft Excel file *transformer_failures_E12_3.xls* and explore its fields.

What is the target variable (the variable that we want to predict)?

2. Close Excel.
3. Launch **SAP Predictive Analytics**
4. Click *Automated Analytics* → *Modeler*

5. Click on *Create a Classification/Regression model*
6. In the Select a Data Source, choose
   a. Data Type: *Excel Files*
   b. Folder: Navigate to the folder where you have saved *transformer_failures_E12_3.xlsx*
   c. Data Set: *Transformer_failures_E12_3.xlsx*
7. **Next**
8. In the Data Description screen, click on *Analyze*
9. The application makes a guess at the types of variables; for instance, *Overloads* is nominal. *Latitude* is continuous.
10. **Next**
11. In the Selecting Variables screen, choose
    a. Target Variables: *Status*
    b. Remove other variables from Target Variables, if any.
12. **Next**
13. In the Summary of Modeling Parameters screen,
    a. Compute Decision Tree: *selected*
    b. Enable Auto-selection: *selected*
14. **Generate**
15. In the *Training the Model* screen, go to Model Overview Report
    a. You see under Selection Process that the Nb. Of Variables Kept is 6. These are the most influencing variables of all the variables that were present in the Excel file
    b. Predictive Power (KI) is 0.4774. This is the model quality or accuracy. It measures the percentage of the target variable that the influencing variables can explain.
    c. Prediction Confidence (KR) is 0.9462. This is the model reliability, robustness. It measures the model's ability to perform to the same level of the training data when new data are presented to it.
    d. **Next**
16. In the *Using the Model* screen,
    a. Click on *Model Overview.* Report Type: Executive Report
    b. See the Profit curve. The red line indicates the random prediction of the target variable.

The blue line is the validation data performance. The green line is the training data performance

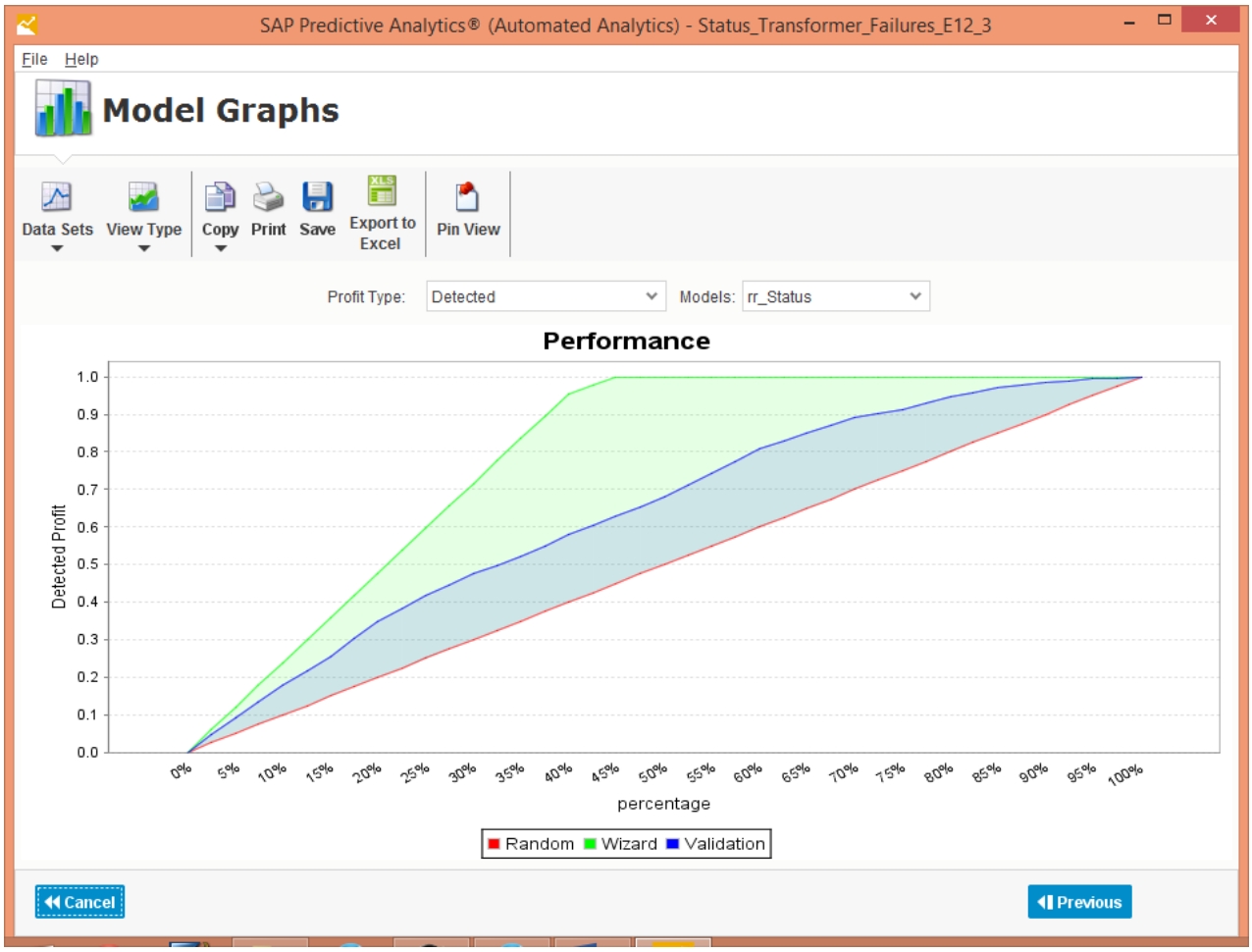

Figure 1: Model Performance

17. Previous

18. In the *Using the Model* screen,

a. Click on *Contributions by Variables*. The chart shows the six most influencing variables. *Overloads* is the most influencing variable for predicting Failure.

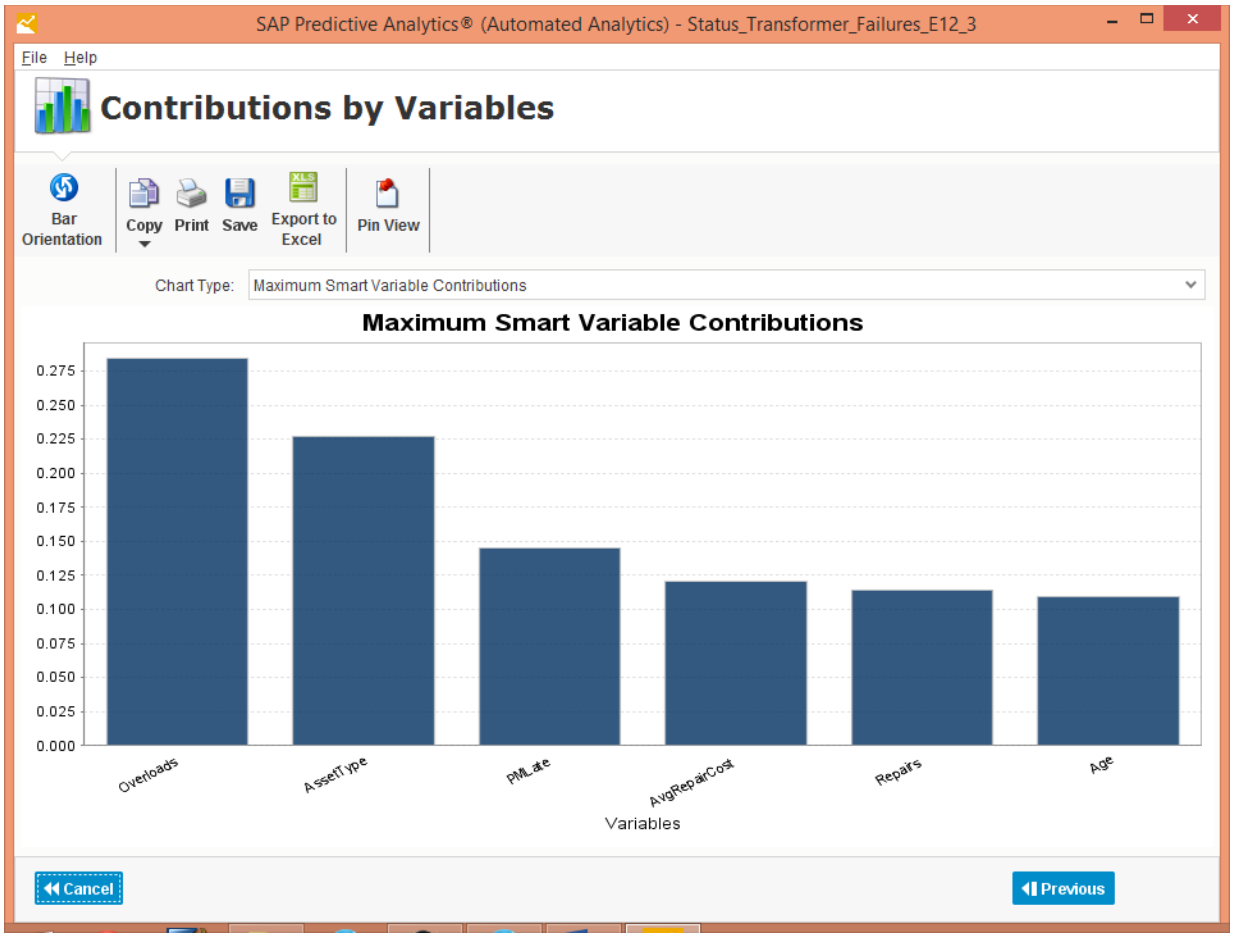

Figure 2: Influencers

19. Previous

20. In the *Using the Model* screen,

a. Click on Confusion Matrix

b. Examine the Confusion Matrix. The matrix shows the rate at which the model predicts true Failure and true Oks, as well as false Failures and false OKs.

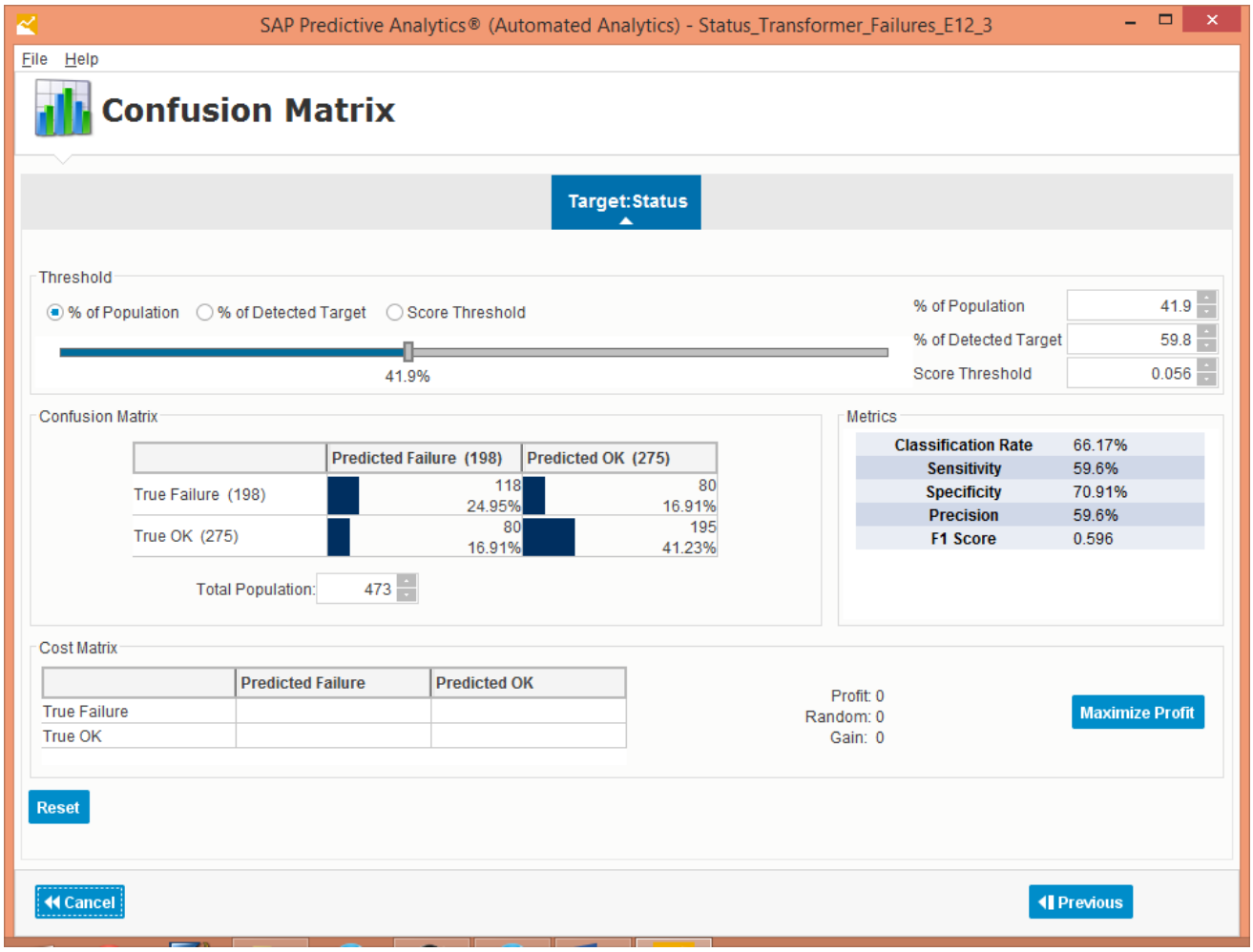

Figure 3: Confusion Matrix

21. Previous

22. In the *Using the Model* screen,

a. Click Decision Tree

b. Explore the Decision Tree by expanding various nodes. You can see the population at each node. The most influencing factor is Overloads. At lower levels, you see the other influencers.

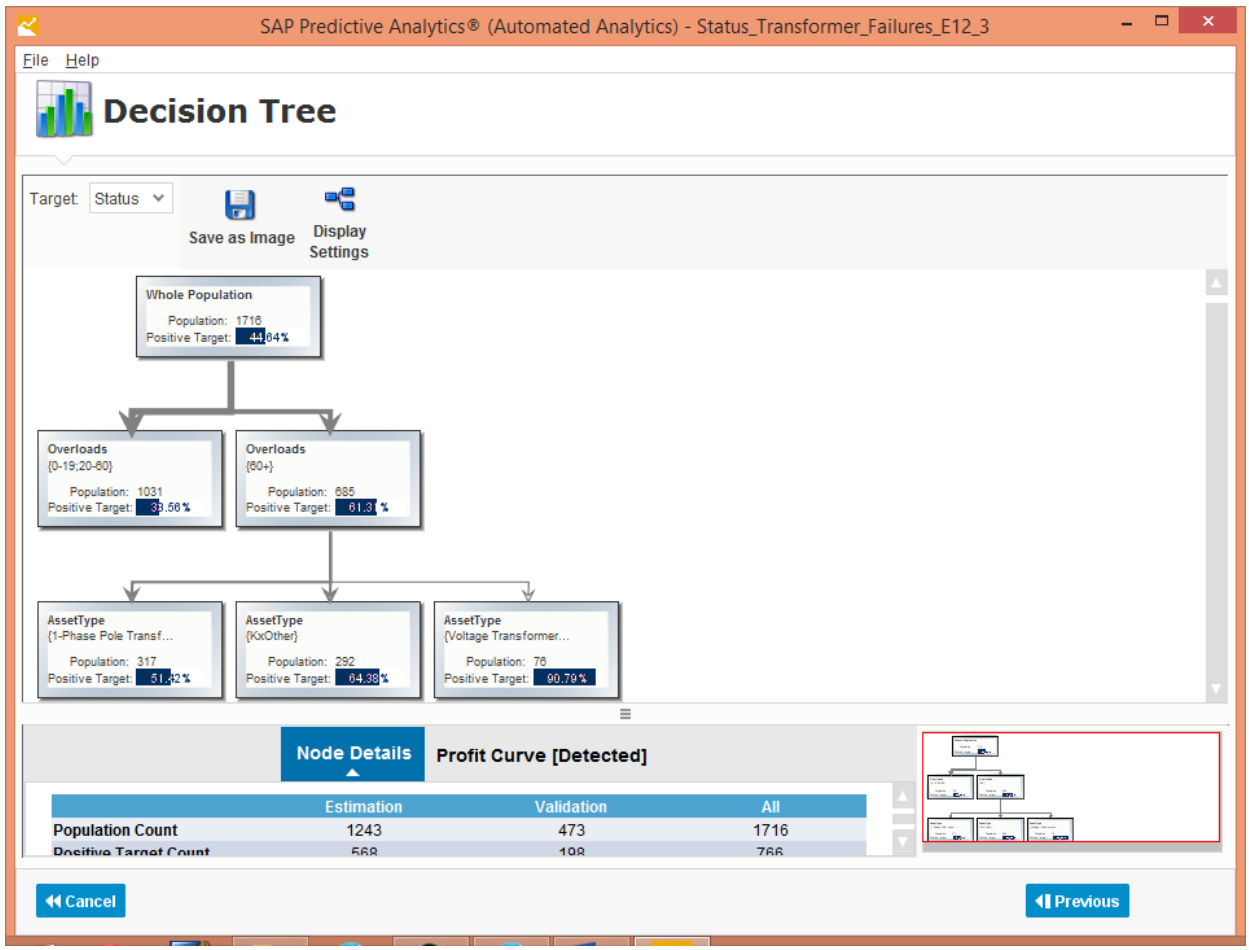

Figure 4: Decision Tree

23. Previous

24. In the *Using the Model* screen,

a. Click on Descriptive Statistics → Date Set Size

b. You can see who many records were used for training (estimation) and how many were used for validation of the data model.

25. Save.

26. To answer the following question, click on Run on the Using the Model screen. Choose Simulation.

Question: Predict the probability of failure of an equipment with the following values:

Overloads*: 0-19,* AssetType*: 1-Phase Pole Transformer,* PMLate*: N,* AvgRepairCost*: 50000,* Repairs*: Original,* Age*: 40*