Chapter 6: Regression Models

- This Chapter demonstrates how to develop logistic regression models for targets with different measurement scales: binary, categorical with more than two categories, ordinal, and continuous (interval-scaled).
- Using an example data set with a binary target, we demonstrate various model selection criteria and model selection methods.
- We also present business applications from the banking industry involving two predictive models: one with a binary target, and one with a continuous target.
- The model with a binary target predicts the probability of response to a mail campaign while the model with a continuous target predicts the increase in deposits that is due to an interest rate increase.
- This chapter also shows how to calculate the lift and capture rates of the models when the target is continuous.

Introduction

- This chapter explores the **Regression** node in detail using two practical business applications—one requiring a model to predict a binary target and the other requiring a model to predict a continuous target.
- Before developing these two models, we present an overview of the types of models that can be developed in the **Regression** node, the theory and rationale behind each type of model, and an explanation of how to set various properties of the **Regression** node to get the desired models.

What Types of Models Can Be Developed Using the Regression Node?

Models with a Binary Target

- When the target is binary, either numeric or character, the **Regression** node estimates a logistic regression.
- The **Regression** node produces SAS code to calculate the probability of the event (such as response or attrition).
- The computation of the probability of the event is done through a *link function*.
- A *link function* shows the relation between the probability of the event and a linear predictor, which is a linear combination of the inputs (explanatory variables).

An Overview of Some Properties of the Regression Node

- A clear understanding of
- **Regression Type,**
- **Link Function,**
- **Selection Model**, and
- **Selection Model Criterion**

properties is essential for developing predictive models.

Regression Type Property

The choices available for this property are Logistic Regression and Linear Regression.

Logistic Regression

- If your target is categorical (binary, ordinal, or nominal), Logistic Regression is the default regression type.
- If the target is binary (either numeric or character), the **Regression** node gives you a Logistic Regression with logit link by default.
- If the target is categorical with more than two categories, and if its measurement scale (referred to as *level* in the Variables table of the **Input Data Source** node) is declared as ordinal, the **Regression** node by default gives you a model based on a cumulative logits link (see Section 6.2.2.)
- If the target has more than two categories, and if its measurement scale is set to Nominal, then the **Regression** node gives, by default, a model with a generalized logits link (see Section 6.2.3).

Linear Regression

- If your target is continuous or if its measurement scale is interval, then the **Regression** node

Link Function Property

- In this section, we discuss various values to which you can set the **Link Function** property and how the **Regression** node calculates the predictions of the target for each value of the **Link Function** property.
- We start with an explanation of the theory behind the logit and probit link functions for a binary target.
- You may skip the theoretical explanation and go directly to the formula that the **Regression** node uses for calculating the target for each specified value of the **Link Function** property.
- *See Chapter 6, Section 6.3 for more information.*

Selection Model Property

- The value of this property determines the model selection method used for selecting the variables for inclusion in a regression, whether it is a logistic or linear regression.
- You can set the value of this property to None, Backward, Forward, or Stepwise.
- If you set the value to None, all inputs (sometimes referred to as *effects*) are included in the model, and there is only one step in the model selection process.

Backward Elimination Method

- To use this method, set the **Selection Model** property to Backward, set the **Use Selection Defaults** property to **No**, and set the **Stay Significance Level** property to a desired threshold level of significance (such as 0.025, 0.05, 0.10, etc.) by opening the Selection Options window.
- Also, set the Maximum Number of Steps to a value such as 100 in the Selection Options window.

Backward Elimination Method When the Target Is Continuous

- The backward elimination method for a continuous target is similar to that for a binary target.
- However, when the target is continuous, an *F* statistic is used instead of Wald Chi-Square statistic for calculating the *p*-values.
- To demonstrate the backward elimination property for a continuous target, we used the process flow shown Display 6.20.

Display 6.20

Business Applications

- The purpose of this section is to illustrate the use of the **Regression** node with two business applications, one requiring a binary target, and the other a continuous target.
- The first application is concerned with identifying the current customers of a hypothetical bank who are most likely to respond to an invitation to sign up for Internet banking that carries a monetary reward with it.
- A pilot study was already conducted in which a sample of customers were sent the invitation.
- The task of the modeler is to use the results of the pilot study to develop a predictive model that can be used to score all customers and send mail to those customers who are most likely to respond (a *binary* target).

Business Applications

- The second application is concerned with identifying customers of a hypothetical bank who are likely to increase their savings deposits by the largest dollar amounts if the bank increases the interest paid to them by a preset number of basis points.
- A sample of customers was tested first. Based on the results of the test, the task of the modeler is to develop a model to use for predicting the amount each customer would increase his savings deposits (a *continuous* target), given the increase in the interest paid.

Remark:

Had the test also included several points representing decrease, increase, and no change from the current rate, then the estimated model could be used to test the price sensitivity in any direction.

However, in the current study, the bank is interested only in the consequences of a rate increase to customer's savings deposits.

Business Applications

- Before using the **Regression** node for modeling, you need to clean the data.
- This very important step includes purging some variables, imputing missing values, transforming the inputs, and checking for spurious correlation between the inputs and the target.
- Spurious correlations arise when input variables are derived in part from the target variables, generated from target variable behavior, or are closely related to the target variable.
- Post-event inputs (inputs recorded after the event being modeled had occurred) might also cause spurious correlations.
- We performed a prior examination of the variables using the **StatExplore** and **MultiPlot** nodes, and excluded some variables (see Chapter 6, Section 6.4).

A modeler

- Only a modeler who knows where his data came from, how the data were constructed, and what each variable represents in terms of the characteristics and behaviors of customers will be able to detect spurious correlations.
- As a precaution before applying the modeling tool, you should examine every variable included in the final model.

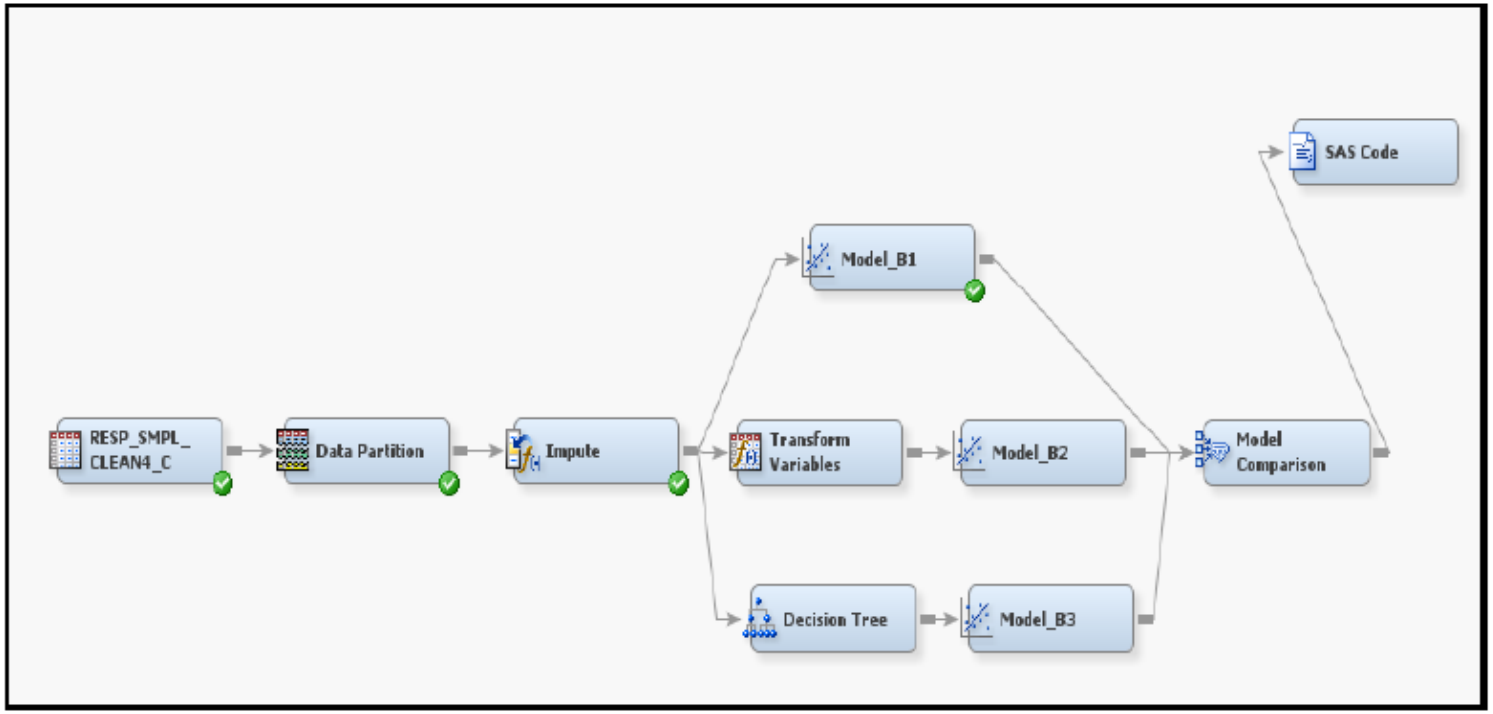
Important Points in Models

- The first model is based on untransformed inputs. This type of model can be used to make a preliminary identification of important inputs.
- For the second model, the transformations are done by the **Transform Variables** node, and for
- The third model transformations are done by the **Decision Tree** node.
- The **Regression** node creates a dummy variable for each category and uses it in the regression equation.
- These dummy variables capture the interactions between different inputs included in the tree, which are potentially very important.

Logistic Regression for Predicting Response to a Mail Campaign (See Section 6.4)

The process flow for each model is shown in Display 6.54.

Display 6.54



Summary - Chapter 6

- Four types of models are demonstrated using the Regression Node. These are models with binary, ordinal, nominal (unordered) and continuous targets.
- For each model, the underlying theory is explained with equations which are verified from the SAS code produced by the Regression Node.
- The Regression Type, Link Function, Selection Model, and Selection Options properties are demonstrated in detail.

Summary

• A number of examples are given to demonstrate alternative methods of model selection and model assessment by setting different values to the Selection Model and Selection Criterion properties. In each case, the produced output is analyzed.

• Various tables produced by the Regression Node and the Model Comparison Node are saved and other tables are produced from them.

Summary

- Two business applications are demonstrated - one requiring a model with binary target, and the other a model with continuous target.

1) The model with binary target is for identifying the current customers of a hypothetical bank who are most likely to respond to a monetarily rewarded invitation to sign up for Internet banking.

2) The model with continuous target is for identifying customers of a hypothetical bank who are likely to increase their savings deposits by the largest dollar amounts if the bank increases the interest paid to them by a preset number of basis points.

3) Three models are developed for each application and the best model is identified.