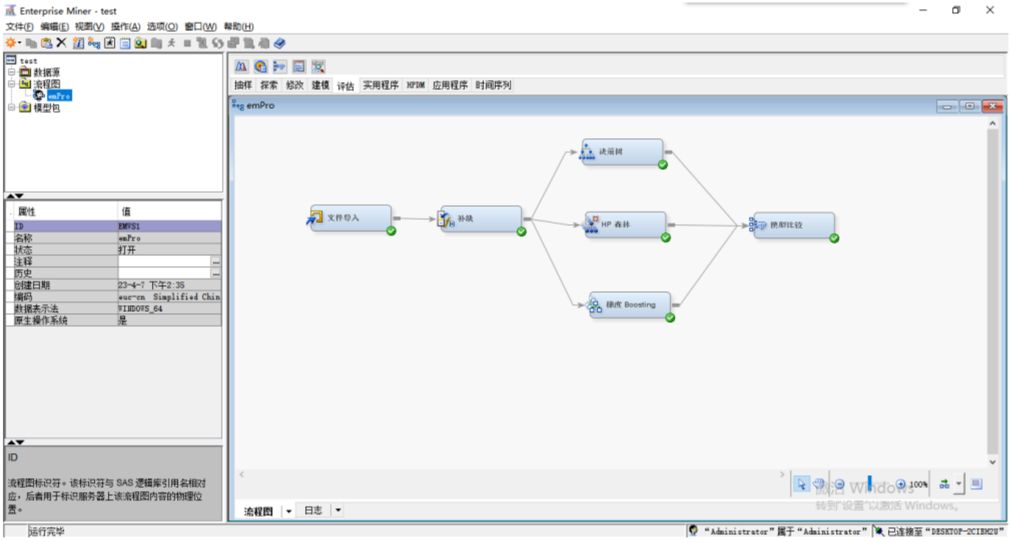
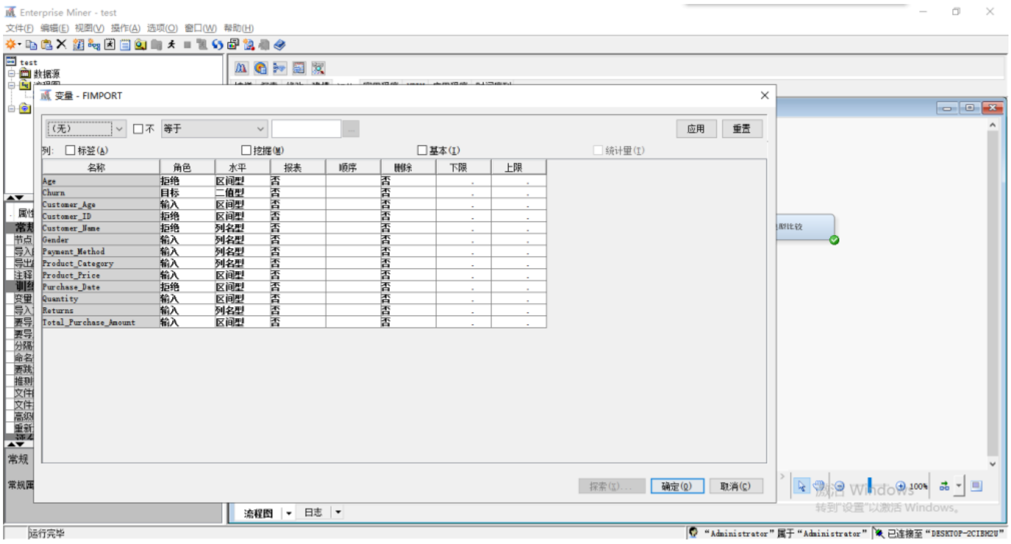
WQD7005 DATA MINING

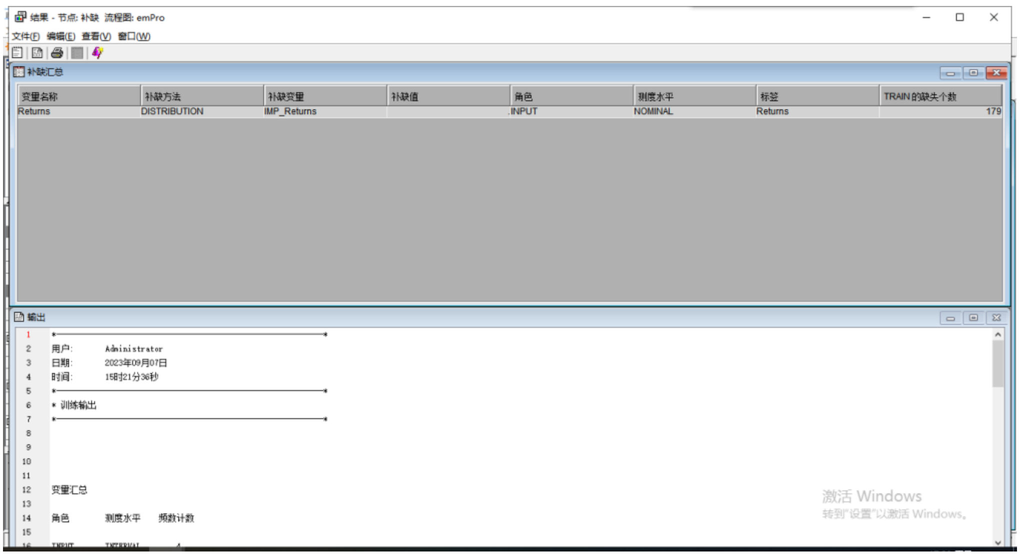
SAS e-Miner：



Data collection and import

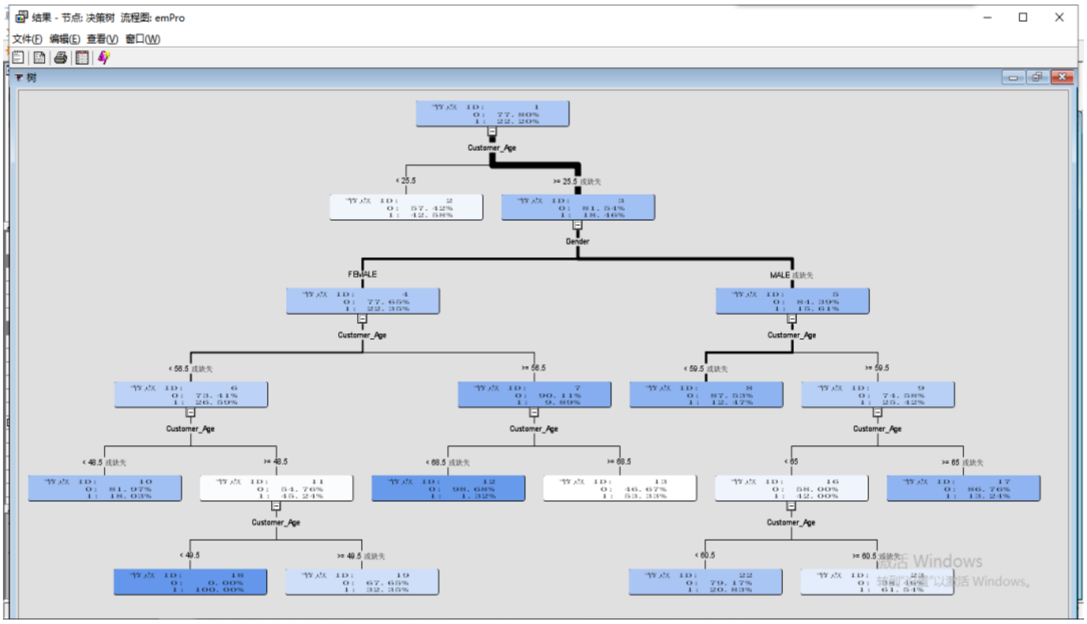


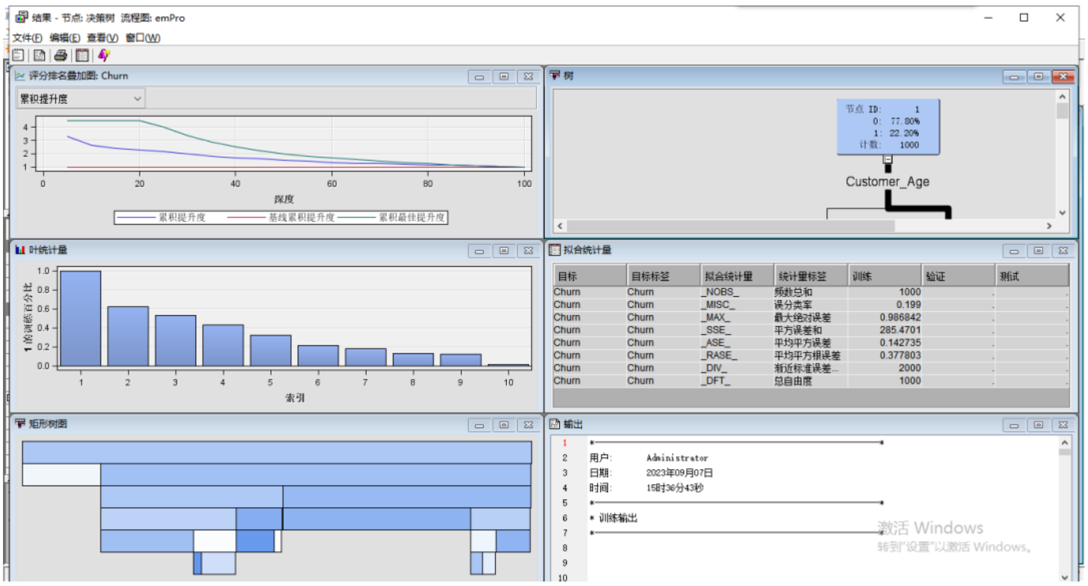
Data cleaning and preprocessing



In the dataset, the dependent variable (target variable) should be "Churn" because this is the outcome I want to predict. All other variables can be used as independent variables (explanatory variables) to predict whether a customer will churn.
Because there are two ages in the data. So I set up a rejection for one of them, because the name, ID, and date cannot be used for modeling, so I also set up a rejection.

Decision Tree：

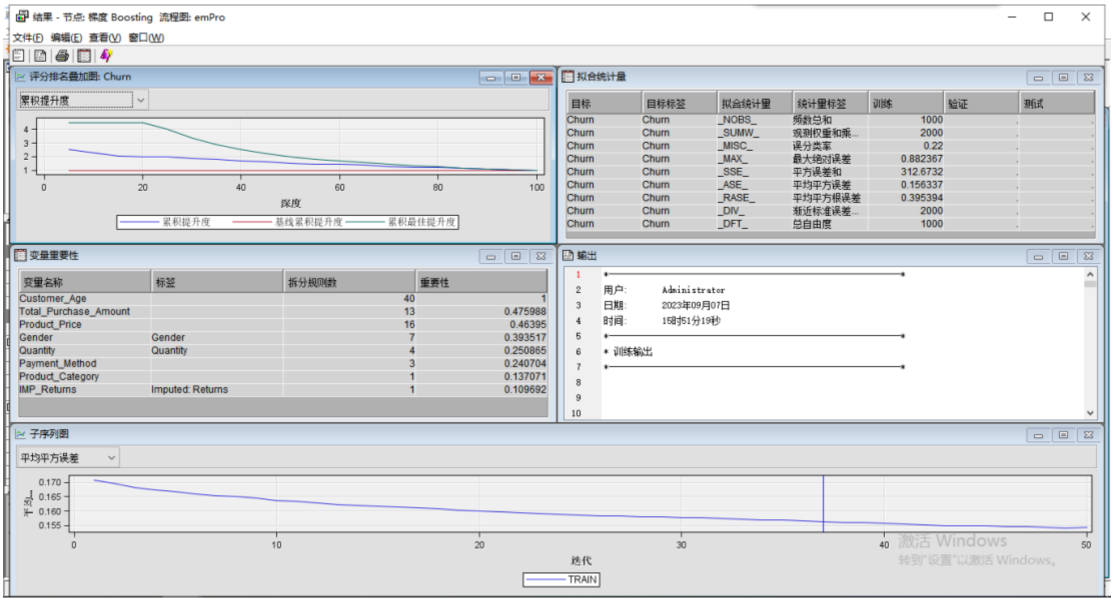




The problem is that an overfitted model relies too much on specific features in the training data and therefore may perform poorly on new, unknown data because those features may not apply to the new data. On the contrary, underfitting models perform poorly on both training and test data because they do not fully learn the characteristics of the training data and do not understand the data well enough. To solve these problems, we can use techniques such as limiting the depth of the decision tree or the minimum number of samples at a node to prevent overfitting, or consider using ensemble methods such as random forest or gradient boosting, which work by combining multiple models. Improve generalization capabilities to achieve better performance on new data. These methods can effectively improve the performance of the model.

—Random Forest

—Boosting



These two models help items attempt to understand customer behavior. Identify which characteristics most affect customer churn, purchasing patterns, etc. At the same time, it can capture the complex nonlinear relationship between features and target variables. If customer behavior is influenced by the interactions between different variables, these methods can effectively model such effects. Given the previous analysis of decision trees and the construction of many unrelated trees and averaging their predictions, Random Forest is not easily overfitting. Boosting also has mechanisms to prevent overfitting, such as scaling down and random boosting, which involve training sub samples of the data.