Intro to Data Science

1. Clustering:  
   First of all, ignore the “Status” variable. Your goal will be to group customers together based on similar characteristics. Assume your goal is to create clusters that will somehow be useful to your employer. The definition of “useful” may be somewhat subjective.  
   A. Create clusters using hierarchical clustering. Use criteria (including the graph under  
   Cluster Summary) to decide of how many clusters would be the “best”. Explain your  
   decision and include a screenshot of your graph. (5 pts)

A screenshot of a data

Description automatically generated

We chose 5 clusters to represent the data the best. We wanted a low number of clusters between 2-10 and 5 seemed to balance lowering standard deviations and separating the data efficiently.

B. Create clusters using k means clustering. Use a range of k and choose which k creates  
the “best” clusters. Explain your decision and include a graphic, (choices include  
“Parallel Coordinate Plots” and a “Scatterplot Matrix”). (5 pts)

For the k-means clustering, we put in a range of 2-12 and JMP calculated 11 to be the “best” number of clusters. Although there are other factors including understandability that JMP does not consider within K-Means.

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

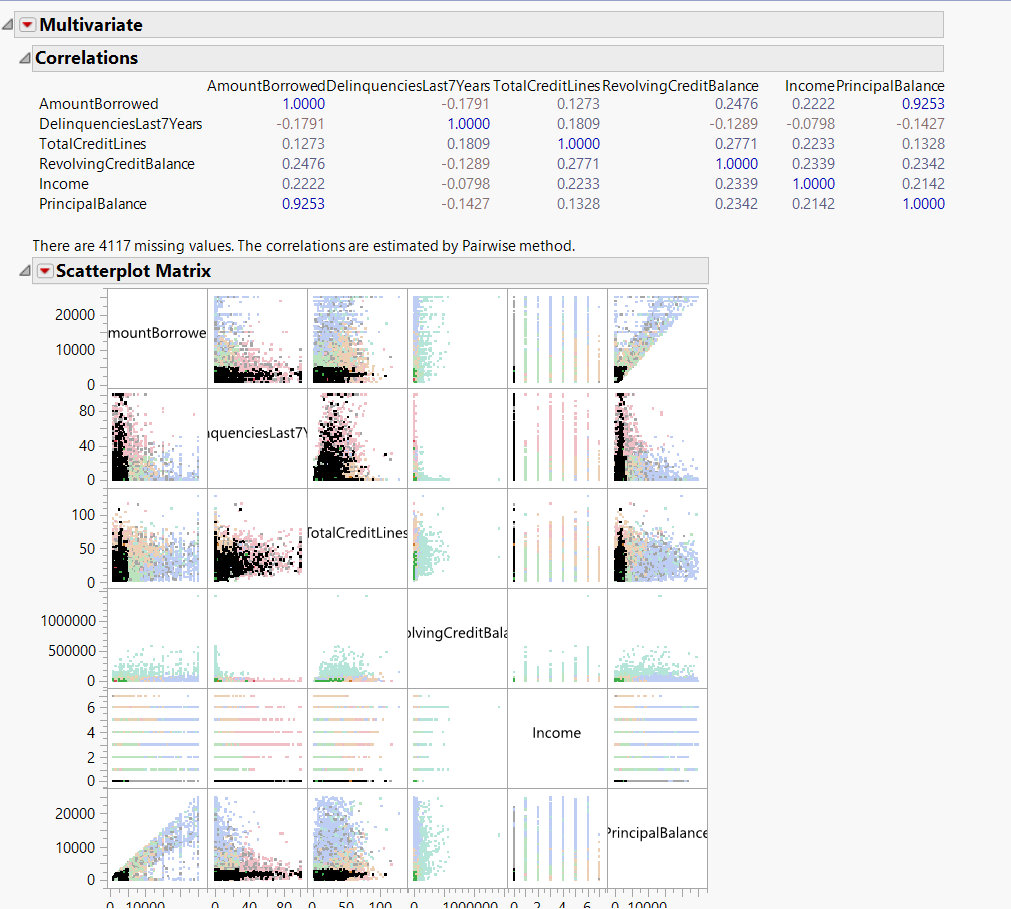
C. Your employer wants to learn something useful about its customers. Write a paragraph  
justifying why one of your clustering results (from either B or C) could be useful to your  
employer. This should include some reference to what the differences are amongst your  
clusters. What can your employer do with this info? (5 pts)

I would say the best number of clusters for an employer would be four. Due to the preexisting categories for credit established: poor, fair, good, excellent. These are simple enough for a customer or employer to understand and doesn’t overcomplicate the data. Although the K-Means clustering says 11 is the best number of clusters, that isn’t easily categorized for presentation. Five clusters is also a good option, but I felt four was easier to color and categorize.

1. Classification

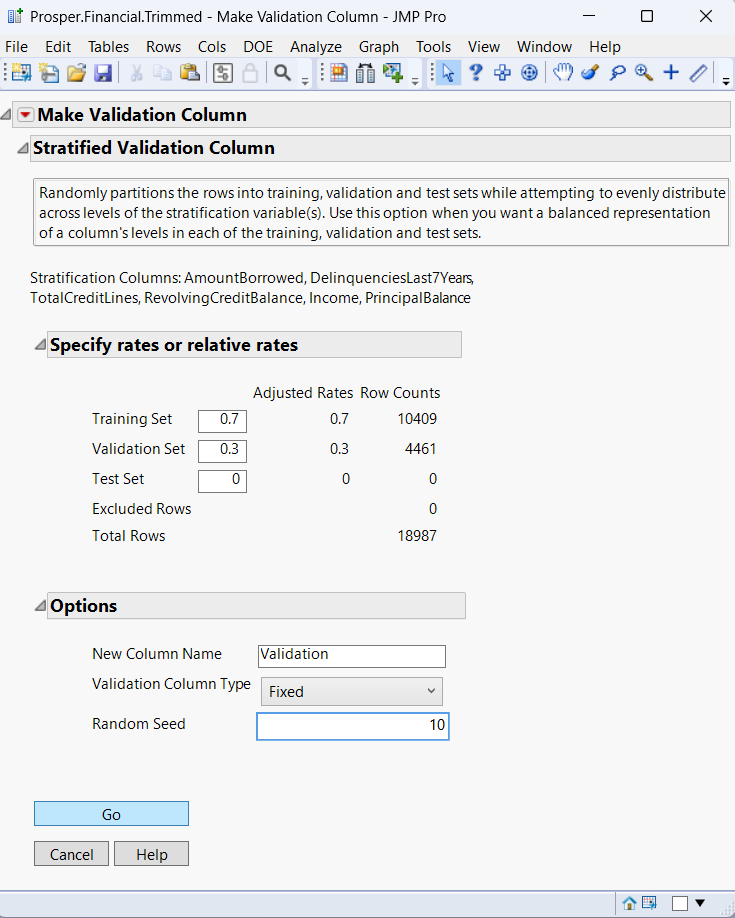
You can also use the same dataset to try to predict “good” and “bad” customers. (We won’t use your clusters from part 1 during this part).

A. Do some exploratory statistics. What variables do you think would be most predictive of “good” and “bad”? Give two graphs or statistics and some short commentary. (5 pts)



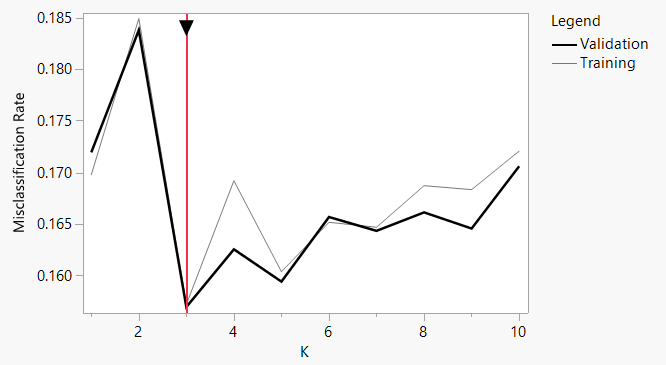
The two variables that will predict ”good” and ”bad” the best would be PrincipalBalance and AmountBorrowed. These two variables have the strongest correlation (0.9253) which indicates that determining on what they are will show whether the client has a ”good” or ”bad” status.

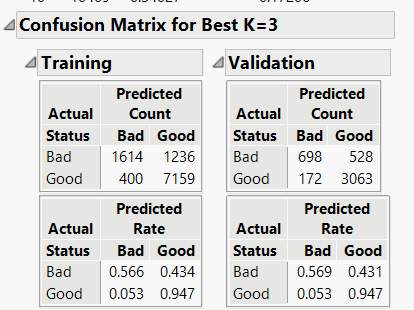
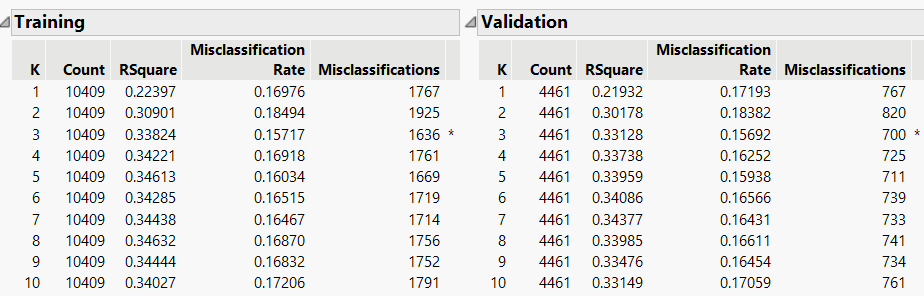
B. Create a 70-30 validation data column. Set seed to 10. (2 pts)



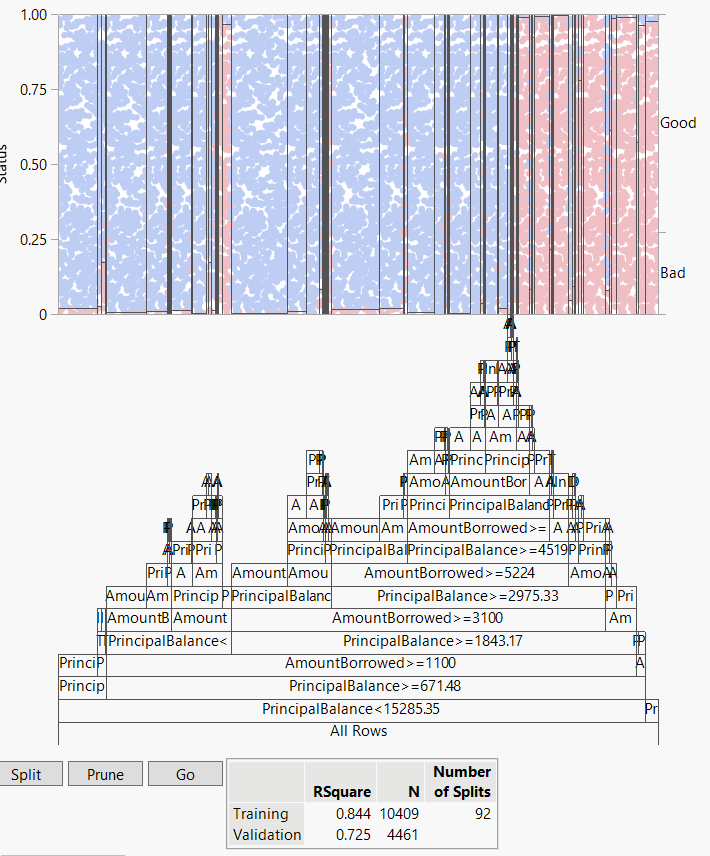
C. Create a couple of different classification models. You can choose two of the following:  
kNN, a basic decision tree, and random forest. For each one, share the misclassification  
rates on your validation set and an image of your confusion matrix. If you make a simple  
decision tree, a screenshot would be good, as well. (5 pts each)

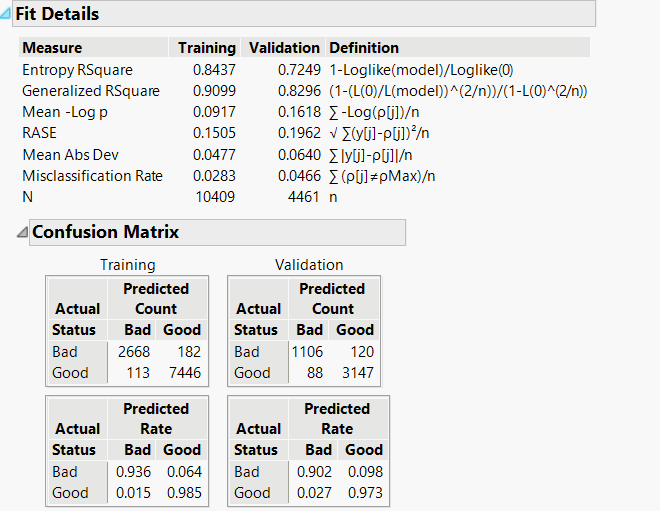
KNN





Decision Tree





D. You should now have created two classification models. Which model would you  
recommend if your goal was accuracy? Which model would you recommend if your goal is  
explainability? How useful do you think your employer would find either model? (5 pts)

From the KNN and Decision Tree classification observed the better one is the Decision Tree. The Decision Tree has a lower misclassification rate for both validation and training while also having a higher R^2 value. Having a lower misclassification rate and higher R^2 will cause for a better and more accurate graph.

1. From the KNN and Decision Tree classification observed the better one is the Decision Tree. The Decision Tree has a lower misclassification rate for both validation and training while also having a higher R^2 value. Having a lower misclassification rate and higher R^2 will cause for a better and more accurate graph.

A. I have taken Heart Disease Dataset from Kaggle which consists of four databases (Cleveland, Hungary, Switzerland, and Long Beach) from 1988. It is a medical dataset used to analyze factors associated with heart disease. This dataset contains 14 columns and 1025 entries. The key variables include:  
 1. age: Patient’s age

2. sex: Patient’s gender (1 = male, 0 = female)

3. cp: Chest pain (categorical, with 4 types of chest pain)

4. trestbps: Resting blood pressure

5. chol: Cholesterol level

6. fbs: Fasting blood sugar

7. restecg: Results of resting electrocardiography

8. thalach: Maximum heart rate achieved

9. exang: Exercise-induced angina (1 = yes, 0 = no)

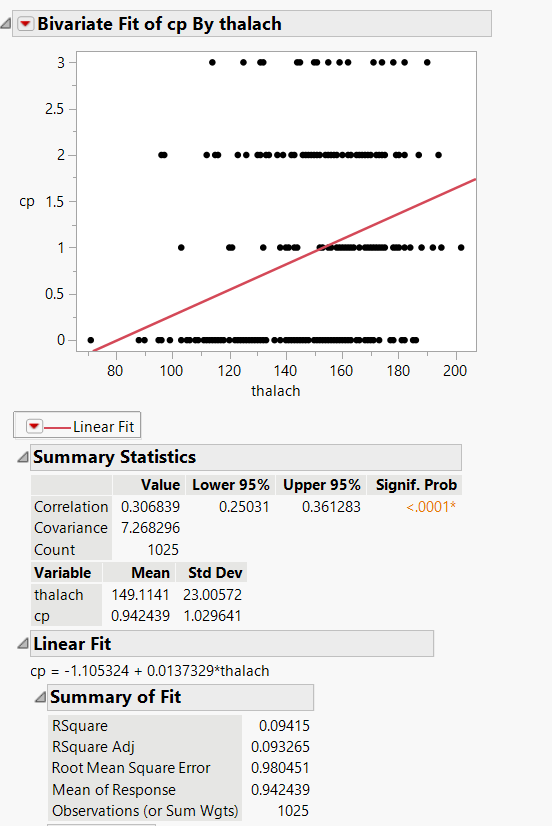
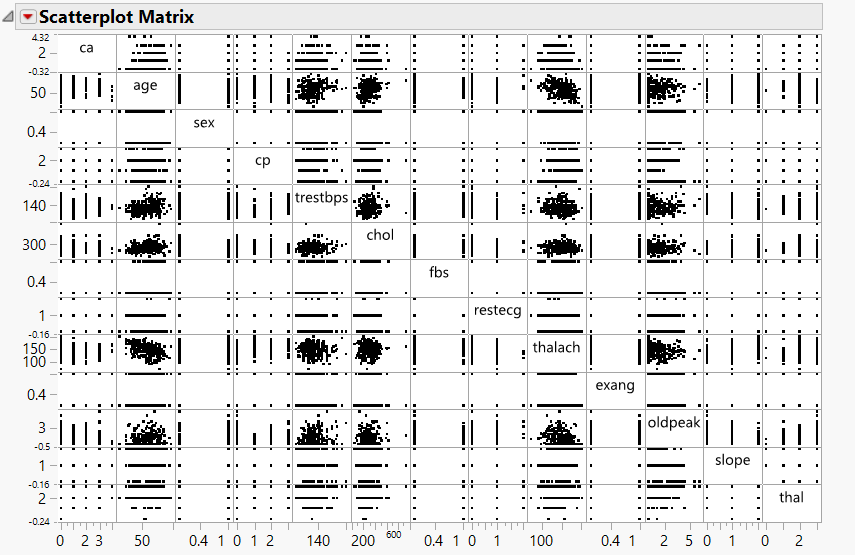
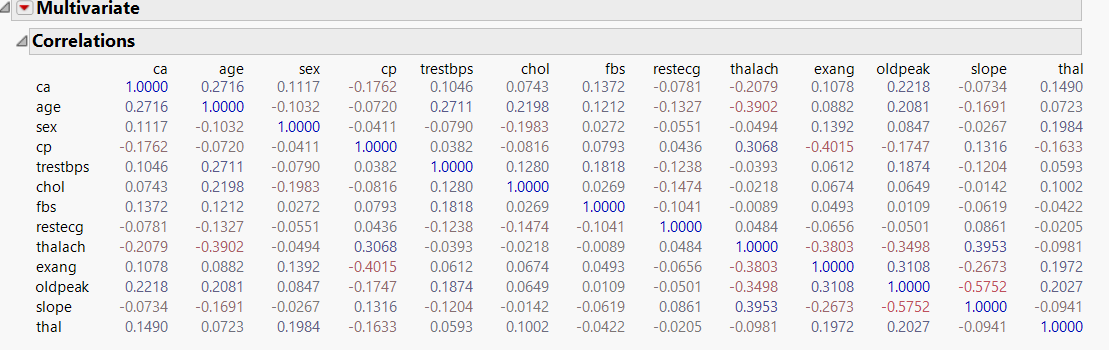
10. oldpeak: Depression induced by exercise relative to rest

11. slope: Slope of peak exercise ST segment

12. ca: Number of major vessels colored by fluoroscopy (0-3)

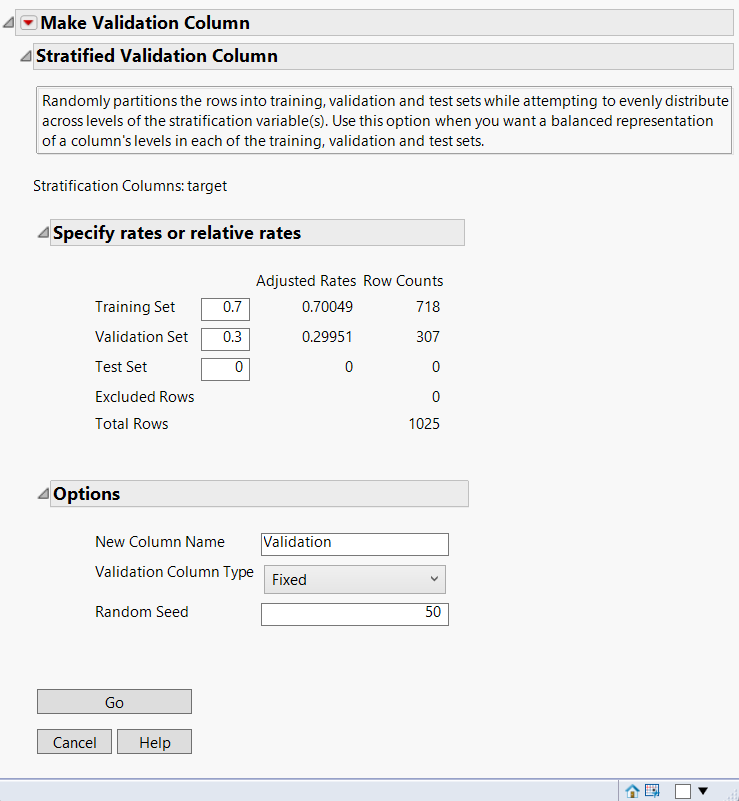
13. thal: A categorical variable indicating thallium stress test results

14. target: Refers to the presence of heart disease (1 = disease, 0 = no disease)

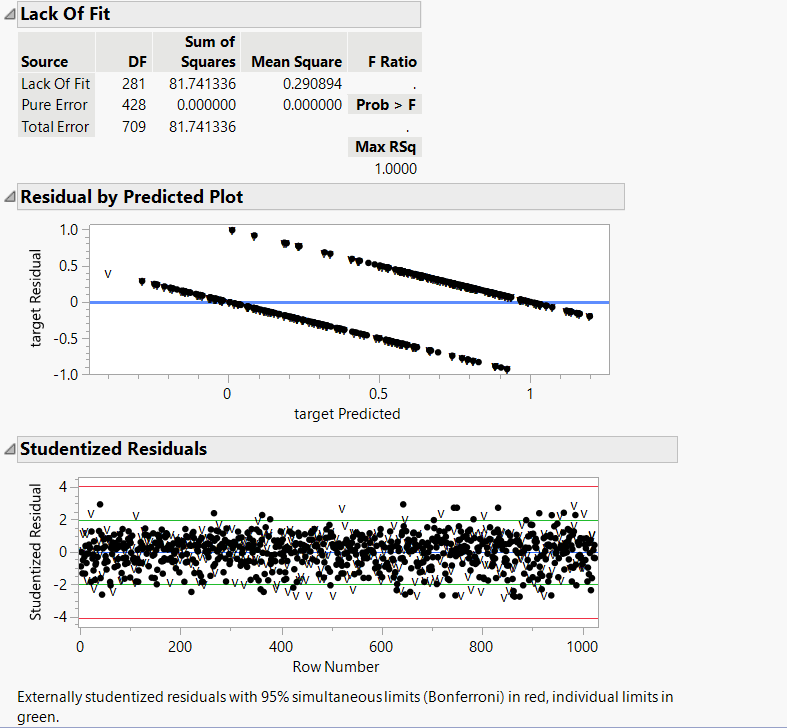
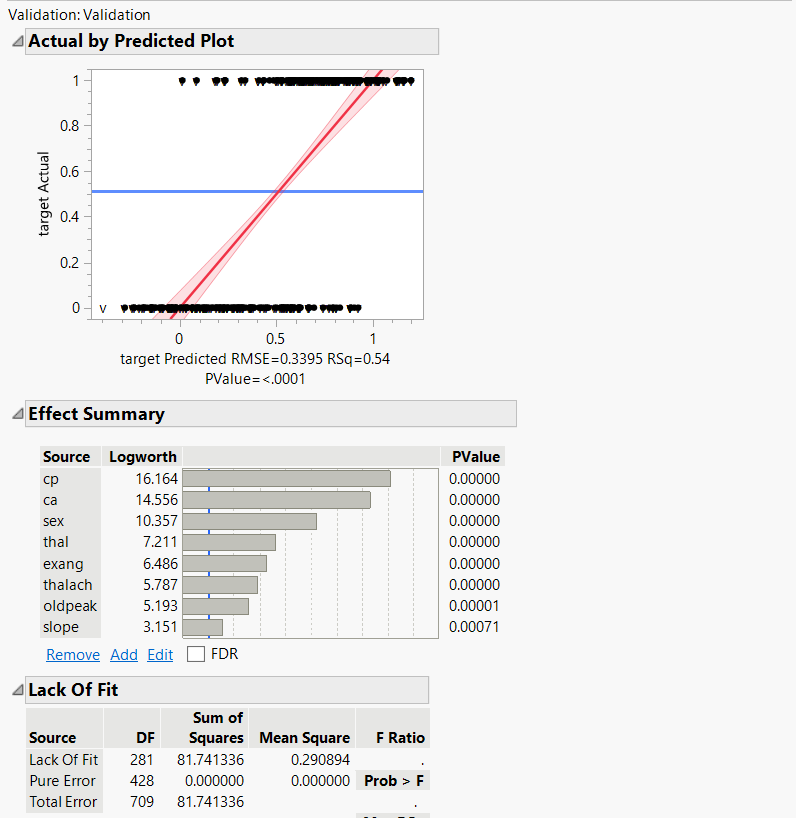
B.

By observing the data cp and thalach seem positively related. Cp and thalach got the strongest correlation. The positive correlation indicates these predictors are useful for modeling.

C. I made a 70 -30 validation column using target as stratification column and seed number is 50.



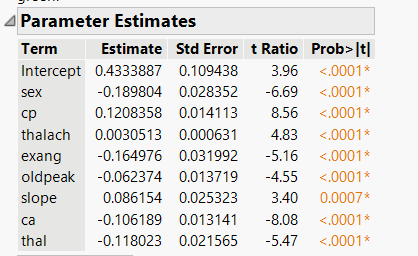
D. I made a regression model using the stepwise algorithm with minimum BIC. I wanted to predict the target column using every other column in this dataset. Here’s what I got from my regression analysis model:



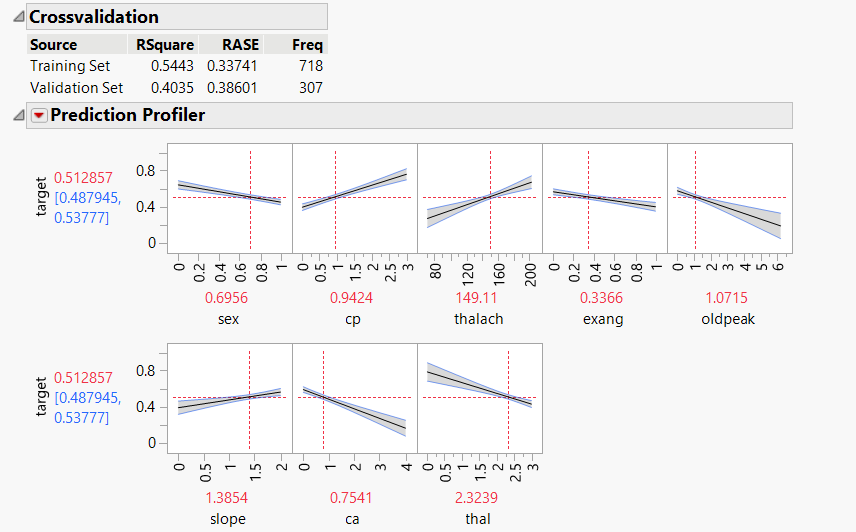
1. p-value is <0.0001, which means that the model is statistically significant.
2. Effect summary suggests that variable like cp, ca and sex have the highest impact, and they are the significant predictors. Logworth and p-values confirm statistical significance for all listed predictors.
3. The regression coefficients for each predictor are shown here. These values describe the relationship between each variable and the target. A negative coefficient indicates an inverse relationship, while a positive coefficient indicates a direct relationship.
4. The training set RSquare is 0.5443, while validation set RSquare is 0.4035. Which indicates that the model generalizes fairly well, although there is a slight drop in performance on unseen data.
5. The validation set’s RASE 0.38601 is somewhat greater than the training set’s RASE 0.33741.
6. There are no notable outliers according to the studentized residuals, which indicates most of the points fall inside the confidence intervals.
7. Summary of Coefficient:

* intercept = 0.4333887
* sex = -0.189804
* cp = 0.1208358
* thalach = 0.0030513
* exang = -0.164976
* oldpeak = -0.062374
* slope = 0.086154
* ca = -0.106189
* thal = -0.118023

E. The coefficients in our model are highly trustworthy. Because all predictors have very small p-values, which indicates strong statistical significance. This means that the relationship between the predictors and the target variable is not due to random chance. High t-ratios confirm that the coefficients are meaningful, and their values are reliable. The estimates are consistent and accurate because standard errors for coefficients are low when compared to their estimates. For example, cp has a coefficient of 0.1208358 and a standard error of 0.014113.



F. The model has a moderately good fit. Training set RSquare is 0.5443, which explains about 54% of the variance in the training data, which is moderately good. Validation set RSquare is 0.4035 on unseen data, which explains about 40% of the variance. Training RASE 0.33741 and Validation RASE is 0.38601. The increase in RASE on validation set is small, meaning that the model generalizes fairly well. Although, its performance slightly decreases on unseen data. There are some overfittings because there is a slight gap between the training and validation RSquare values, but not drastic, so overfitting is moderate.



G. I think the model is highly explainable and useful. Because there is clear identification of significant predictors and their effects on the target variable. Effect Summary and Parameter Estimates show that predictors like cp, ca and sex have the largest impact on the target variable. Patients with particular forms of chest discomfort may have increased target probability because Cp has a beneficial effect on the target. Important clinical links are shown by the detrimental effects of sex and calcium on the target. It is a useful tool for prediction since it can account for more than 40% of the variance in unseen validation data.