CMSE 381 Final Project

Jim Stratton

CMSE 381, Spring 2021

1. Introduction

Across the United States, millions of people use some form of illicit drug at some point in their life. The exact number of users changes depending on what type of drug and what age group is in question. The point of this project is to create a model to predict the age of a user based on a number of predictors such as blood pressure, cholesterol, drug type, and more. Several prediction methods will be used to see which method produces the best result. These methods include Linear regression, subset selection, and tree-based modeling.

2. Related Work

This data set has been used for multiple analysis by various people though www.kaggle.com [1]. It was developed to aid in teaching machine learning techniques to those who are looking for more practice, or those new to the topic. The goal of dataset was to predict type of drug used by an individual based on variables like age, sex, and blood pressure. While the main focus of the previous analyses attempted to predict a different variable, the approach for this project is generally the same.

Previous studies have mostly used a tree-based model to make their drug prediction. For reference we will look at a submission from Kaggle user AkashSDas [2]. For their project, they attained the best parameters for their model and used a decision tree for their analysis. They ended up creating a model that was 98% accurate.

For this project, we will use multiple models and see which one outperforms the others, and that model will be the selected method to make our final prediction. This model is attempting to predict a person's age instead of the drug type used. I have selected Linear regression, subset selection, and decision tree.

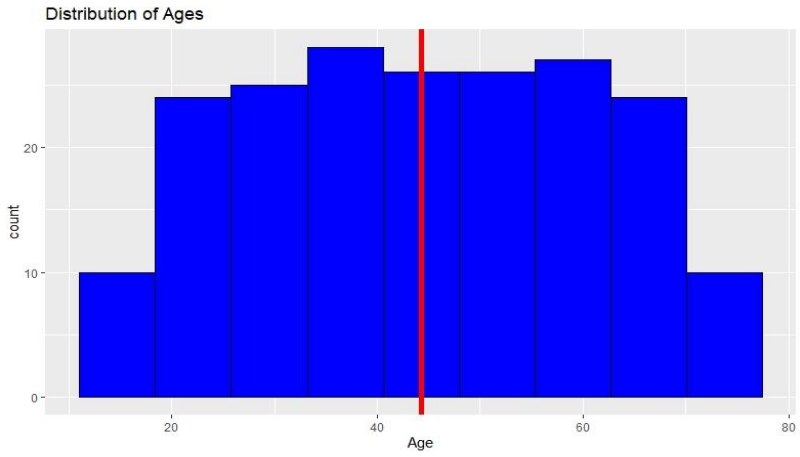
3. The Data

The data set used for this project contains 200 rows of information with 6 columns. A break down of all the predictors with a brief explanation of what they mean can be found below.

Age	Subject's age (int)
Sex	Subject's Sex (M or F)
Blood Pressure (BP)	Blood Pressure Level (High, Low, or Normal)
Cholesterol	Subject's Cholesterol Level (High, Low, or Normal)
Sodium to Potassium (Na to K)	Ratio of Sodium to Potassium (float)
Drug Type	Type of drug (A, B, C, X, or Y)

Admittedly this dataset is a bit small compared sets used for similar projects, but this should be sufficient for our needs. Most variables in the set were categorical by default, so they needed to be converted to numbers that could be used for calculation and model creation. After getting all the variables in a form that could be used, it was split into training and data sets with 80% of the data going to training the model and the remaining 20% going towards testing the model.

To the right we can see the distribution of the variable we are trying to predict. The ages are more or less normally distributed, and the mean is highlighted in red, estimated to be about 44 years old



4. Methods

The first method applied to this dataset was liner regression. Linear regression was used to get some preliminary information. Following Linear regression, Subset selection and tree based modeling were applied to create models. Those models were then compared to see which performed the best.

Linear Regression

The Linear regression model proved to be pretty straightforward. After creating the model from the training set and getting predictions from the test set, an mean squared error (MSE) was calculated. The MSE for this model ended up being 226.76, which does not indicate that linear regression would be an appropriate model to use for this data.

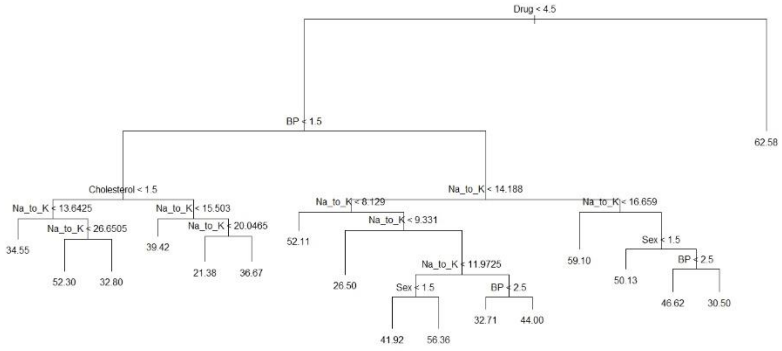
Subset Selection

In order to determine the best route for subset selection, three methods were tested. Getting a summary for forward stepwise, backward stepwise, and best subset methods determined no difference in the number of influential predictors. This is not too surprising considering how few predictors the dataset had to begin with. With that said, the best subset model was selected arbitrarily to go forward.

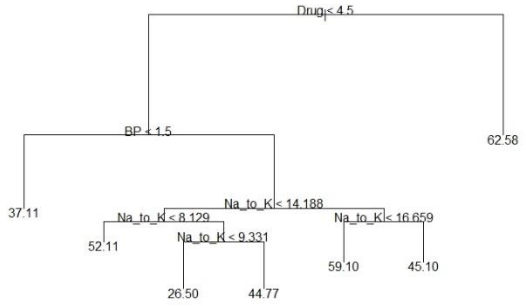
After creating the model, the summary was used to find the best number of predictors needed to make sufficient predictors. Using BIC and R^2 we would only need to use one predictor, however I felt it would be better to have more so the number 2 was chosen. Predictions were made and an MSE was attained. The MSE for best subset was 231.74, which is actually worse than the linear regression model.

Tree Based Model

The final method used in this project was the tree based model. The initial tree produced can be found below. As you can see the tree has many branches and could use some improvement, so some work needed to be done.



In order to improve the model, cross validation was used to determine what size tree would minimize the deviance of the model which ended up being a size of 5. With this new information, the tree was pruned and produced the tree shown on the right.



Unfortunately, I was not able to get an error value in which to quantify this model due to my own errors in R.

5. Results

This project tested three models to find which method would be best for predicting the age of a subject based on a number of factors. After each method was tested, some form of error value was obtained so that their accuracy could be compared. This comparison is illustrated in the following table.

Method	Mean Squared Error
Linear Regression	226.76
Best-Subset Selection	231.74
Tree-Based	NaN

As mentioned before, my own shortcoming in R prevented me from finding a problem which was preventing me from getting an error value for the tree based method. Only the remaining two methods are available to make a choice from, therefore Linear regression performed the best out of the methods tested here.

6. Conclusion

In conclusion, some mistakes were made which prevented a full analysis of the dataset. It is for this reason that I would say the project referenced in section 2 is the superior analysis. If I were to revisit this project, I would definitely change a few things. I think using a larger dataset may make the results more interesting. Having more predictors would give the subset selection a chance to actually have an impact on what information is being sent to the model. I would also be sure to give more time to troubleshooting so that I could attain an error value for each method. In the end, I definitely learned a few things and improved on other skills I knew I needed to work on.

References

1) Tripathi, Pratham. “Drug Classification.” *Kaggle*, 14 Aug. 2020, www.kaggle.com/prathamtripathi/drug-classification.
2) Akashsdas. “Drugs Classifier Using Decision Tree.” *Kaggle*, Kaggle, 1 Apr. 2021, www.kaggle.com/akashsdas/drugs-classifier-using-decision-tree.