Machine Learning Assignment 2 Report

1. Introduction

The telecommunication company has noticed a lot of churns of the costumers, and wants to make a strategy regarding this problem to prevent other costumers leaving. With the given set of information about the costumers, the strategy should be made.

2. Descriptive analysis of data and use of the data

The information given in the document consider the information about costumer's demography (columns: gender, tenure, SeniorCitizen, Partner, Dependents), his/her accounts (columns: Contract, PaperlessBilling, PaymentMethod, MonthlyCharges and TotalCharges) and type of services he/she receives (columns: PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTv, StremingMovies) which makes 20 columns that have information about the costumer, and we have one, additional important column with the information did the costumer churn.

Before working with the document, I have noticed that some of the cells in the table are empty, so I have excluded them from the analysis and also changed the value of the rows where there were two different values with the same meaning (for example No and No internet service).

I have also mapped the SeniorCitizen values from 0 and 1 to No and Yes, and changed the tenure into the classification order, placing the data in groups:

1-12 Months-> one year

12-24 Months-> two years

24+-> Old costumer

Also, I have excluded the information I don't need, CostumerID and tenure (which is now in the groups), and changed the churn words to left and stayed so the plots are more easily read.

3. Explanation of the used packages

In this project, I have used the following packages:

Rpart (Recursive Partitioning and Regression Trees) -recursive partitioning for classification, regression and survival trees. I have used this package for the decision tree representation.

Plyr(Tools for Splitting, Applying and Combining Data)- a set of tools that solves a common set of problems: you need to break a big problem down into manageable pieces, operate on each piece and then put all the pieces back together. For example, you might want to fit a model to each spatial location or time point in your study, summarise data by panels or collapse high-dimensional arrays to simpler summary statistics. I have used this package to map some values into the new values in the information.

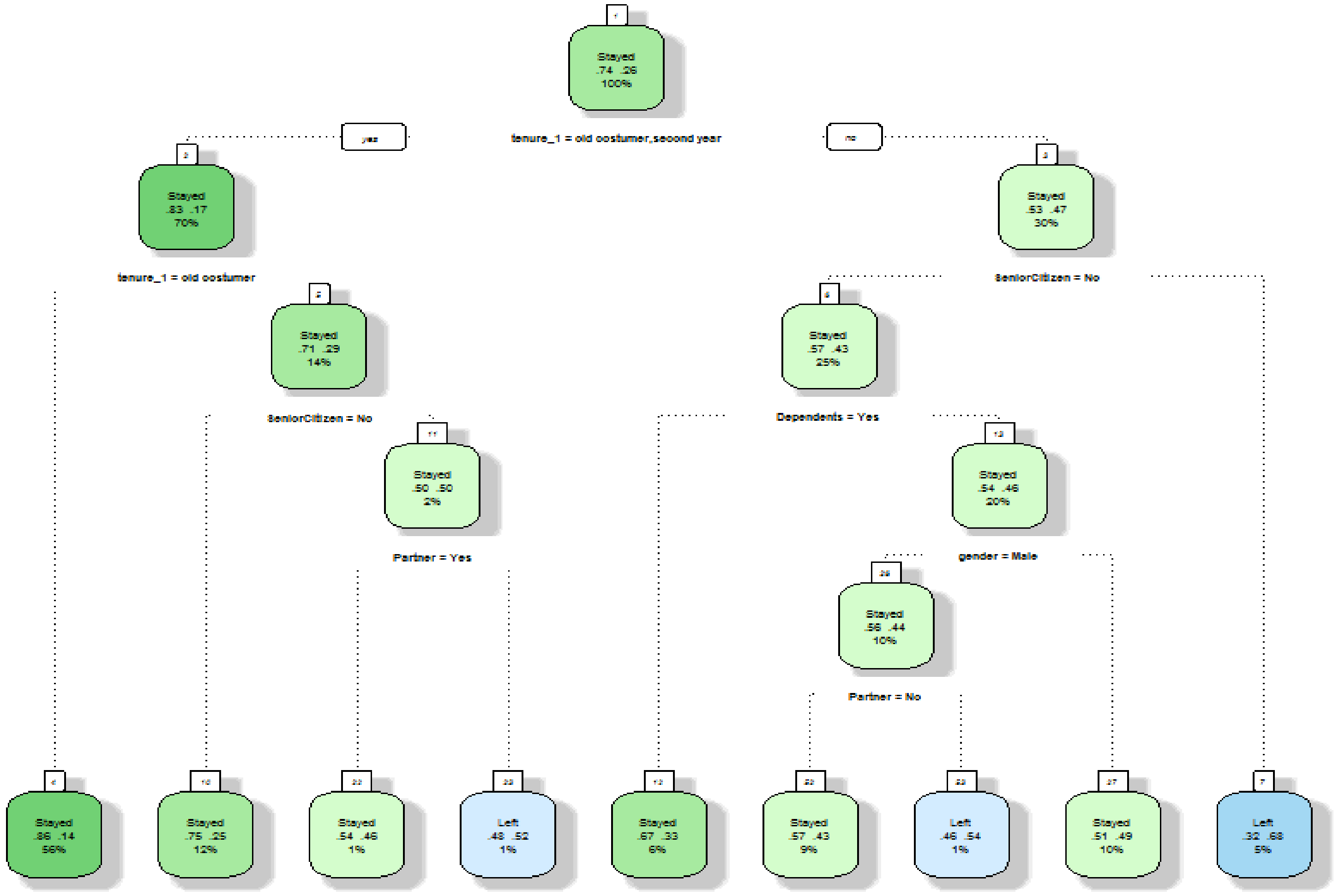
Rpart.plot, RColorBrewer, rattle and caret- used for plotting the results of the tree.

4. Steps and interpretation of the output

After I have finished the optimization of the data with the steps in the 2. , I have divided the data into the training and testing, giving training 80% of the information (5636 rows) and testing 20% of the information (1406 rows). Using the training data, I make 3 decision trees, based on the ground of the information, so I can get the most important types of services that don't function, demographic information that is important and information about the account that has to be improved. I divide data this way because I think the strategy should address the biggest issues in all of the categories.

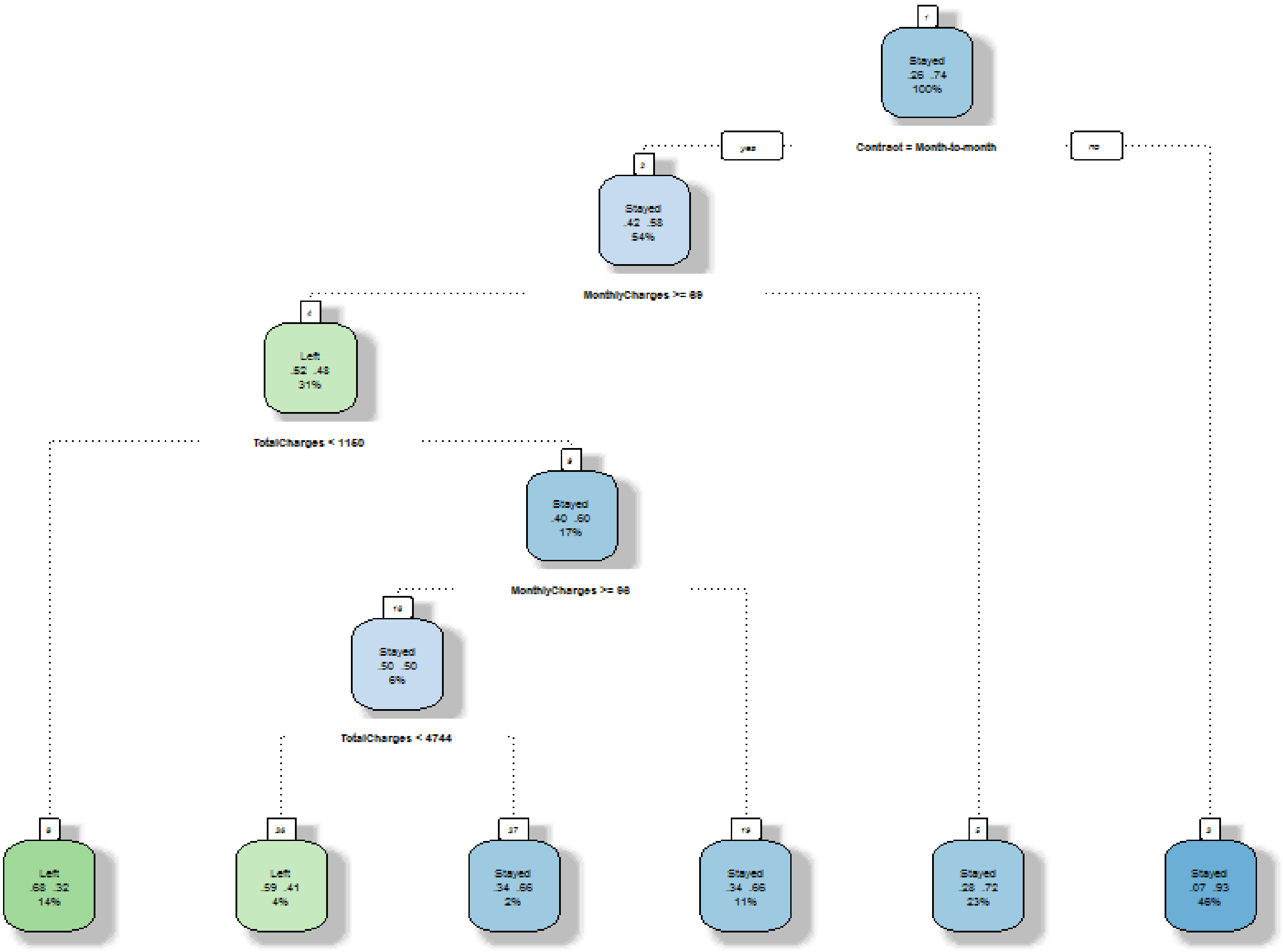
I get the results that the most important reasons are:

Demographic info: Most of the people who left were the senior citizen in their first year of tenure.



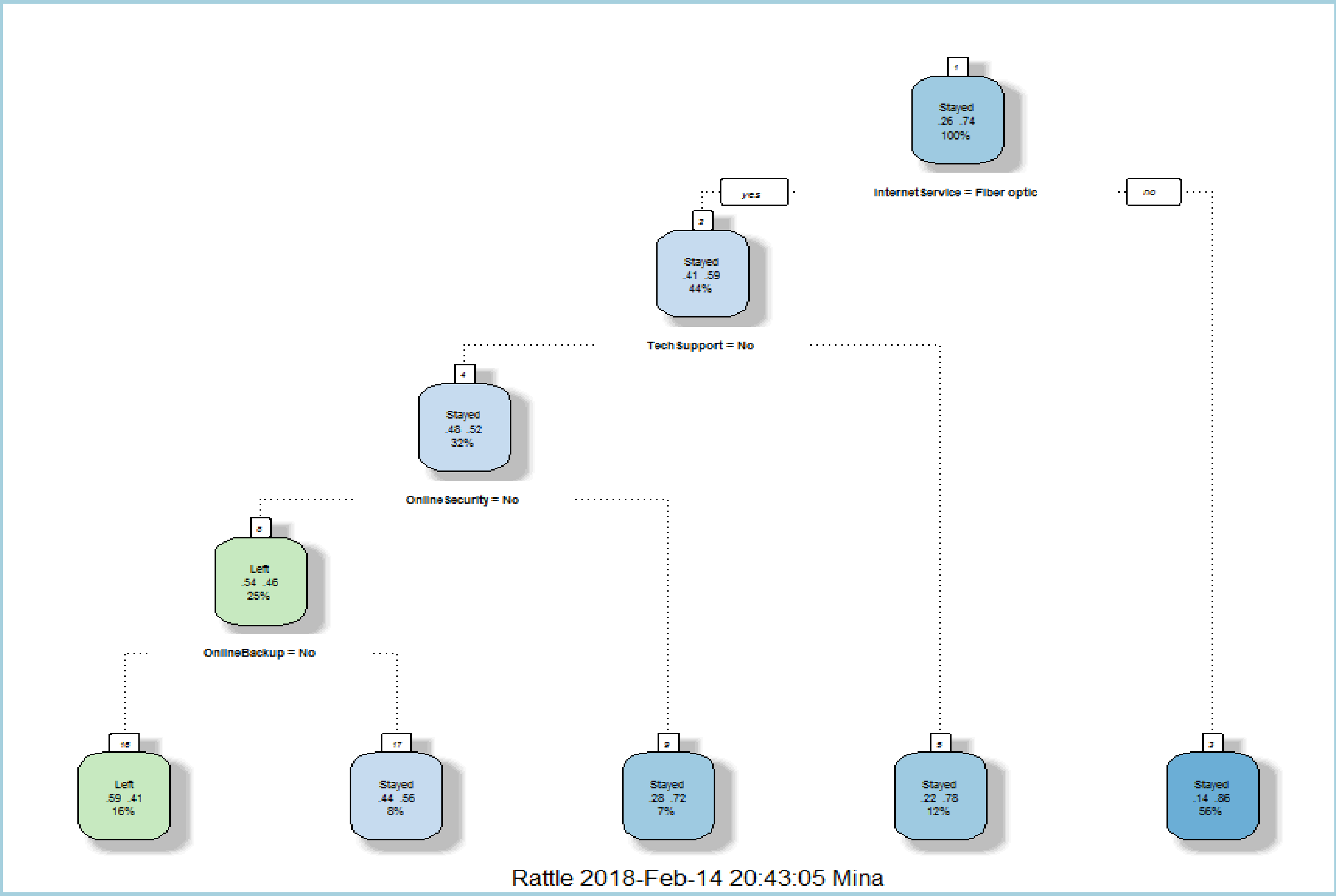
Rattle 2018-Feb-14 20:16:38 Mina

Account information: I first get a very big tree, showing the most important features are Contract, MonthlyCharges and TotalCharges. Then I make the tree with only these three features to see the addiction between them better, concluding that most of the costumers who left were the ones with month to month contract, with monthly charges over 69.

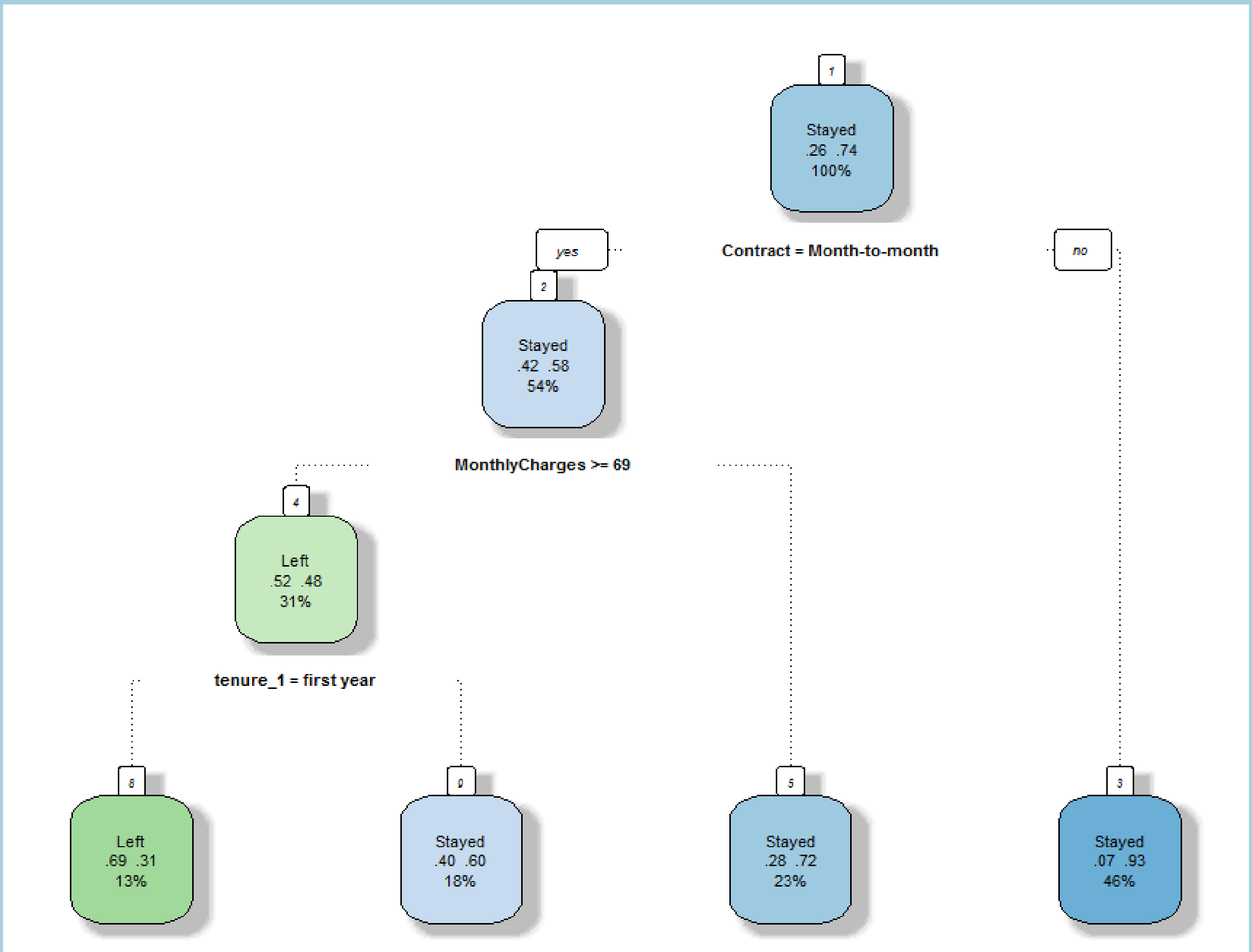


Rattle 2018-Feb-14 20:38:52 Mina

Type of service: Using the similar method of first finding the most important features, and then using them, I concluded that costumers who had fiber optics, without online security, online backup and tech support were the ones leaving the most.



When I compare these results, using the methods like before and using the new decision tree, I get:



Rattle 2018-Feb-14 20:45:37 Mina

Which shows the most important features are contract, monthly charges and tenure.

5. Strategy

Given the results from the last tree, it is obvious that the most costumers leaving are the ones who are in the first year of tenure, with monthly charges over the 69, with the contract month to month.

Company has to address this issue first, so the special offer can be offered to the costumers in the first year of tenure, or the option month to month can be excluded from the offer, preventing the costumers to leave so soon. Also, discount can be offered for the first few months if the contract is longer than one year for example, which can prevent costumers of leaving in the first months.

Even know the issue of fabric option optics internet option isn't on this tree, special attention should be given to the costumers using this feature.

6. Evaluation

I have compared the accuracy of the whole tree with the testing data and got that that decision tree works with 70% accuracy, which is the average score of the decision trees. When a company makes the strategy, the biggest problems have to be solved first, so using decision tree for this problem is ok, because all the important issues will be presented and considered. If the training set of data is bigger, decision tree can be more accurate.

7. Conclusion

Decision tree approach is a good method, with ok accuracy, and it shows diagrams that can easily be read, and for the problem like in this assignment, they are a good solution, because the most important features can easily be seen and used for making the strategy.