**D209 Performance Assessment II**

Renzo Espinoza

College of Information Technology, Western Governors University

D209: Data Mining

Dr. Festus Elleh

October 24, 2022

**A1: Proposal of Question**

The research question that I will attempt to answer in my analysis is: using decision trees, can one predict the tenure of a customer accurately?

**A2: Defined Goal**

In the telecommunications industry, it costs 10 times more to obtain new customers than to retain existing ones. For this scenario, the telecommunications company I am performing the analysis for is interested in retaining their customers, so the ‘Tenure’ field in the dataset is the ideal variable to center my analysis on. The goal of the analysis is to identify the features of a customer that influence their tenure the most, in hopes of allowing the company to target those features in their plans for maximizing customer retention. The dataset contains many features of patients that I could potentially use as predictors.

**B1: Explanation of Prediction Method**

The prediction method I chose, regression trees, analyzes the data by performing linear regression, but instead of having one model for the entire dataset, the data is conditionally split into partitions with simpler models corresponding to each partition. This prediction method is useful when analyzing categorical and continuous features with nonlinear and complex relationships to a target variable. I believe this method is suitable for this scenario considering that there are many categorical and continuous variables in the dataset with a weak linear relationship to ‘Tenure’.

**B2: Summary of Method Assumption**

Since the prediction method is non-parametric, it does not make distribution assumptions of the data. According to the United States Environmental Protection Agency (2022), decision trees do not assume any particular form of relationship between the independent and dependent variables. Decision trees work well for this scenario, because the dependent variable has a skewed distribution and has very few linear relationships with other variables.

**B3: Packages or Libraries List**

|  |  |
| --- | --- |
| **Library** | **Usage** |
| Numpy and Pandas | Data storage and manipulation |
| Matplotlib and Seaborn | Data visualization |
| Scikit-learn | Implement decision tree, split data into training and testing sets, perform cross validation and hyperparameter tuning, and compute the mean squared error and R-squared of each model |

**C1: Data Preprocessing**

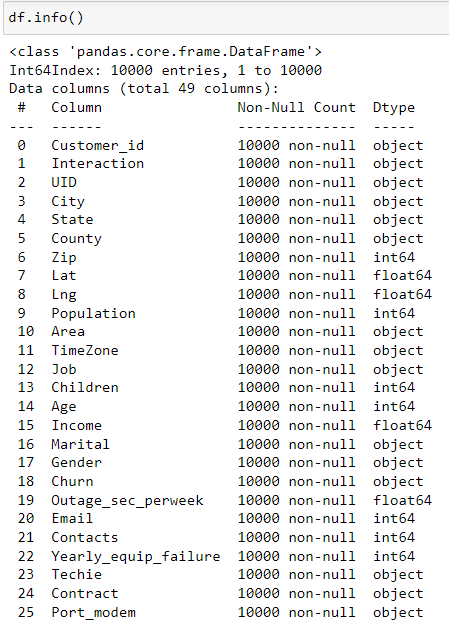
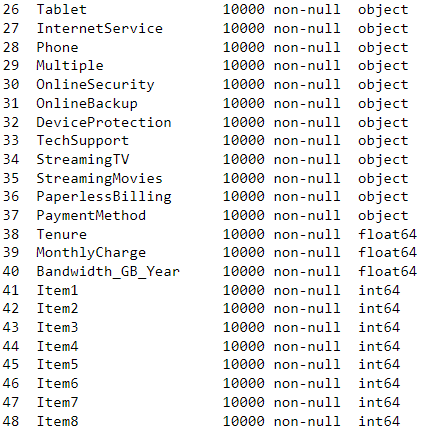
One data preprocessing goal that I have is to encode categorical data. Decision trees accept both categorical and continuous data, but Scikit-learn’s ‘DecisionTreeRegressor’ requires all independent variables to be in numerical format.

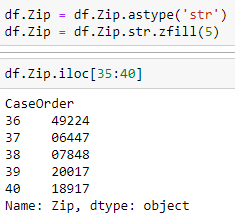
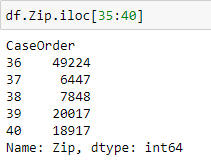
**C2: Data Set Variables**

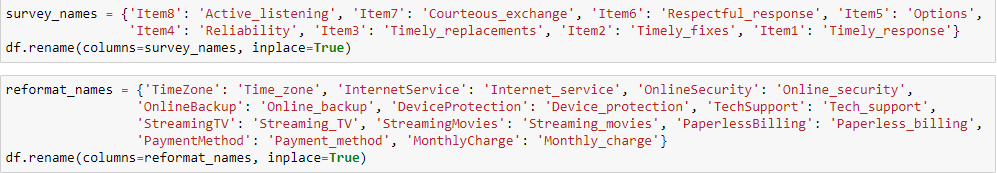
|  |  |
| --- | --- |
| **Variable** | **Type** |
| Tenure | Continuous |
| Bandwidth\_GB\_Year | Continuous |
| Device\_protection | Categorical |
| Options | Categorical |
| Online\_backup | Categorical |

**C3: Steps for Analysis**

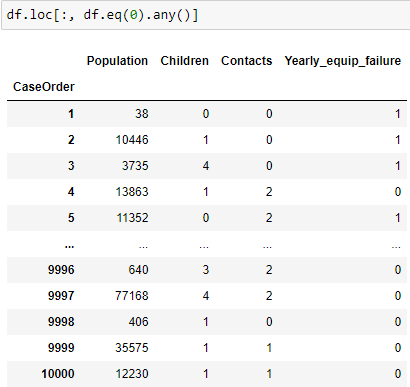
To prepare the data for analysis, I started off by using the ‘Dataframe.info’ method from the Pandas library to get a bird’s-eye view of the data that I imported into a Pandas Dataframe.

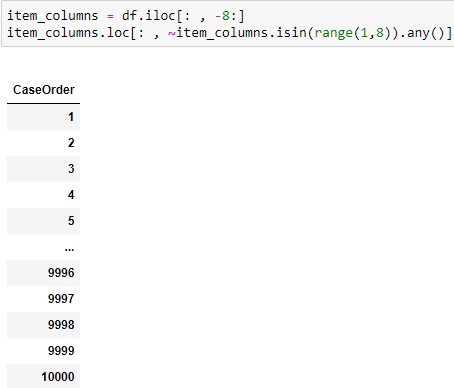


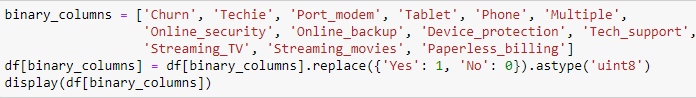
From there I see that the categorical field ‘Zip’ is being imported as an integer. To fix this, I converted the column to String and added the missing leading zeros.

Next, I want to rename certain columns that don’t fit the standard naming conventions of the other columns. I also renamed the survey columns to more easily identify what each survey item pertains to.

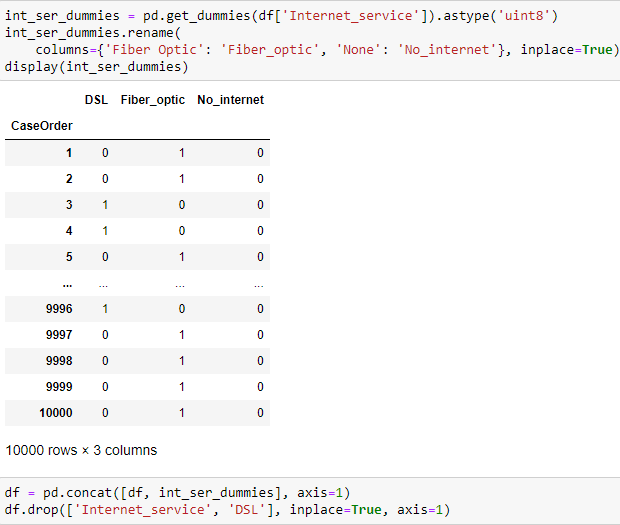
According to the summary of the dataset there are no nulls found in the data, but it is still possible for data to be missing. I queried the dataset for 0’s using the ‘Dataframe.loc’ method and the results showed that there are zero-values inside the ‘Population’ column. The rest of the columns in the results do not imply an input error.



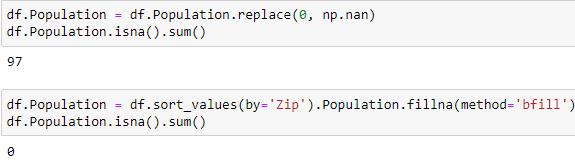
**** I also queried the ‘Item’ columns for values outside of the 1-8 range that the surveys range from. Thankfully, no errors appeared in the eight columns.

Next, I want to convert several categorical binomial variables to integers, represent ‘Yes’ and ‘No’ with 1 and 0, respectively.

I will then perform one-hot encoding on suitable categorical variables. I used the ‘get\_dummies’ method from the Pandas library to create the dummy variables, I then concatenate dummy variables to the dataset and drop the original variable and one of the dummy variables to use as my reference group. Below I show the code for one-hot encoding the column ‘Internet\_service’, I repeat this process for ‘Gender’, ‘Area’, and ‘Contract’.

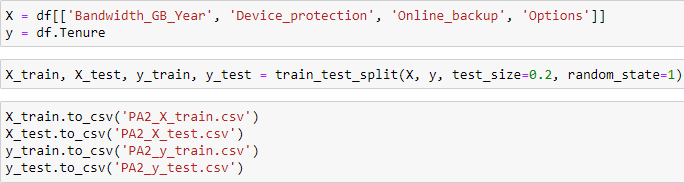


Finally, I want to impute the 0 values found in the ‘Population’ column. To do this I performed backwards fill after sorting the dataset by the ‘Zip’ column.

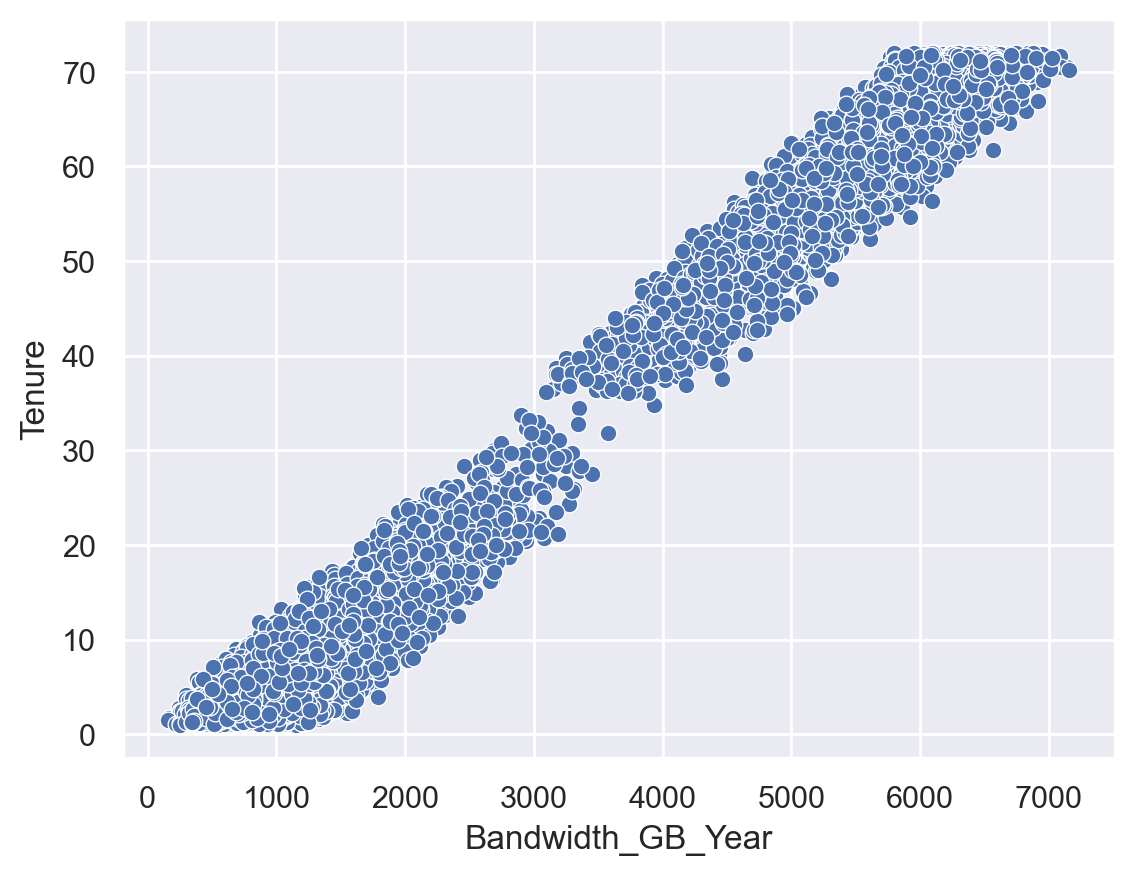


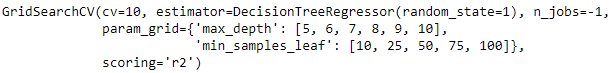
**C4: Cleaned Data Set**

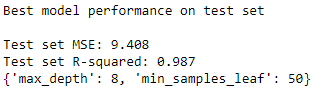
The cleaned dataset will be attached to my submission under the name ‘PA2\_cleaned\_data.csv’.

**D1: Splitting the Data**

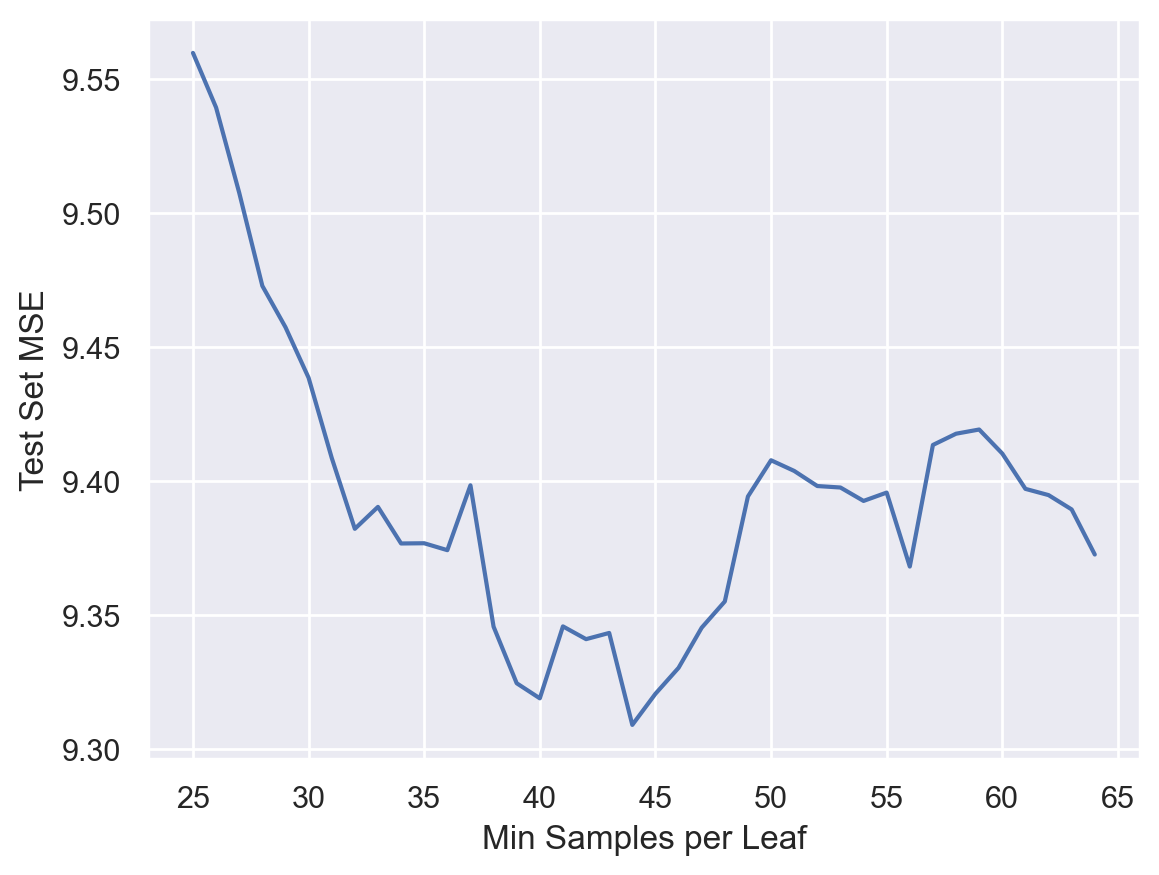
**D2: Output and Intermediate Calculations**

 After cleaning the data, I explored it to find relationships with the target variable ‘Tenure’. The most promising feature by far is ‘Bandwidth\_GB\_Year’ with a correlation coefficient of 99%. After that there is a considerable drop off in terms of correlation, the next three features that I will use for the decision tree are ‘Device\_protection’, ‘Options’, and ‘Online\_backup’ and they each have correlation coefficients ranging from 2-3%.

 Using the Scikit-learn library, I instantiated the regression tree using ‘DecisionTreeRegressor’. To determine the optimal parameters for the tree, I used ‘GridSearchCV’ to search through two different sets of values for the ‘max\_depth’ and ‘min\_samples\_leaf’ parameters with ten-fold cross validation and R-squared as the accuracy metric. I gradually lowered the minimum amount of samples per leaf until the accuracy no longer improved. This process helped me narrow down the range of values for each parameter.

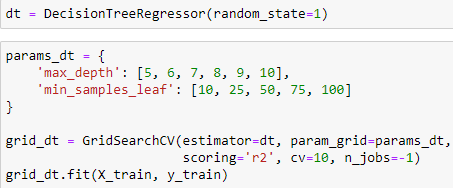


Next, I plotted mean squared error of the regression tree’s predictions on the testing set at different values of ‘min\_samples\_leaf’. The plot shows that a value of 44 results in the lowest MSE for the test set, the maximum depth of the tree is set to 8. Obviously going lower would improve the MSE for the training set but to avoid over-fitting I will use these values for my tree parameters.

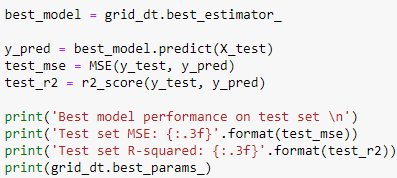


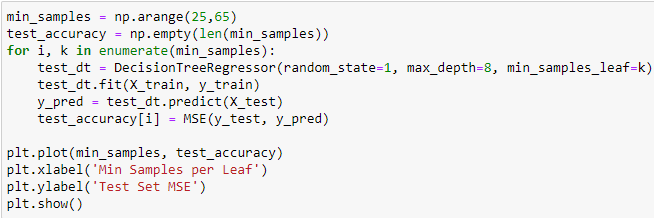
**D3: Code Execution**

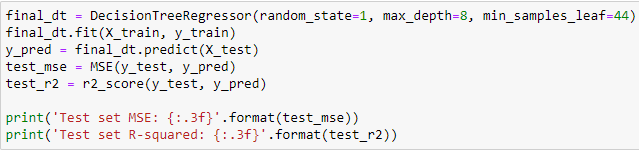
Instantiating decision tree and performing hyperparameter tuning:



Printing the results:

****

****Plotting the test set MSE at different values of ‘min\_samples\_leaf’:

****Final regression tree with MSE and accuracy:

**E1: Accuracy and MSE**

The accuracy, or R-squared, of the regression tree when fitted to the testing split is 98.7% when the minimum samples per leaf is set to 44 and the maximum tree depth is 8. The MSE of the decision tree is 9.31. These metrics suggests that the decision tree is very accurate and explains most of the variance in the data.

**E2: Results and Implications**

While the decision tree performs well, the resulting model is very complex and hard to understand. Aside from ‘Bandwidth\_GB\_Year’, there are no independent variables that are even somewhat related to the dependent variable ‘Tenure’. A simple linear regression would likely outperform this tree while being much easier to interpret and explain. This analysis at least implies that the bandwidth that a customer uses per year is very strongly correlated with the tenure of a customer.

**E3: Limitation**

One limitation of my analysis is that I used just one single tree. According to Bruce et al. (2020), random forest and boosted tree algorithms outperform a single decision tree nearly every time.

**E4: Course of Action**

Since customers who use a lot of bandwidth every year are highly likely to stay with the company longer, my recommended course of action would be to target heavy internet bandwidth users. I would suggest targeted advertisements towards users of streaming services and gamers. Offering promotions that provide faster or better value internet could attract those type of potential customers.

**F: Panopto Recording**

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=1fcc4f9b-65b9-4ae4-863b-af39011364a9

**G & H: Sources**

Bruce, P., Bruce, A. G., & Gedeck, P. (2020, May). Practical statistics for data scientists: 50+ essential concepts using r and python. O'Reilly.

United States Environmental Protection Agency. (2022, February 3). *Basic Analyses*. EPA. Retrieved October 21, 2022, from https://www.epa.gov/caddis-vol4/basic-analyses-1