Noah Cina

7/10/2024

D209 – Machine Learning

Performance Assessment Task 2

***Part I: Research Question***

*A.  Describe the purpose of this data mining report by doing the following:*

*1.  Propose****one****question relevant to a real-world organizational situation that you will answer using****one****of the following prediction methods:*

*•   decision trees*

*•   random forests*

*•   advanced regression (i.e., lasso or ridge regression)*

Given the current dataset, can average customer bandwidth usage per year be accurately predicted with a random forest regression model?

*2.  Define****one****goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.*

The goal of this data analysis is to determine which customer variables in the dataset are accurate predictors of “Bandwidth\_GB\_Year” and to produce a model that can predict the average yearly bandwidth consumption of new customers based on those variables with acceptable accuracy.

***Part II: Method Justification***

*B.  Explain the reasons for your chosen prediction method from part A1 by doing the following:*

*1.  Explain how the prediction method you chose analyzes the selected data set. Include expected outcomes.*

Random forests use decision trees as a base model and make many different trees by sampling random entries in the dataset without replacement. The benefit of using many small decision trees is that it introduces randomness into the model building process and can benefit from the aggregation of many smaller models to produce one large model that performs better than one decision tree.

*2.  Summarize****one****assumption of the chosen prediction method.*

Random forests assume that the dataset is large enough to grow sufficiently diverse trees. The benefit of random forests is most apparent when dealing with a large amount of data, as the diversity of trees is what allows them to make more accurate predictions than single decision trees.

*3.  List the packages or libraries you have chosen for Python or R and justify how*each*item on the list supports the analysis.*

Scikit-learn – robust modelling package for python that includes the model to be used and supports splitting data into training and test sets, performing cross validation, and assessing model validation statistics.

Pandas – standard data science library for python. Useful for working with tables as dataframes and performing operations on whole datasets.

Missingno – used to visualize missing data.

Scipy – used to perform correlation statistics on the dataset for determining which variables were appropriate to use in the classification model. Used to generate random numbers for the random search cross validation.

***Part III: Data Preparation***

*C.  Perform data preparation for the chosen data set by doing the following:*

*1.  Describe****one****data preprocessing goal relevant to the prediction method from part A1.*

Random forests require that categorical variables be expressed as binary decisions, so categorical variables used in the model will need to be re-expressed as dummy variables using one-hot encoding. All new categories will be kept, as dropping one category to prevent multicollinearity does not apply to this model.

*2.  Identify the initial data set variables that you will use to perform the analysis for the prediction question from part A1 and group*each*variable as numeric or categorical.*

The variables used for the regression model are as follows:

Categorical:

Churn

OnlineSecurity

OnlineBackup

StreamingTV

StreamingMovies

Contract\_Two Year

InternetService\_DSL

InternetService\_Fiber Optic

InternetService\_None

Numeric:

Tenure

MonthlyCharge

*3.  Explain the steps used to prepare the data for the analysis. Identify the code segment for*each*step.*

All relevant code is contained in the block labelled “#Part C3 Data Preparation” in the attached .ipynb file.

The dataset was loaded into a variable and checked for missing values using missingno.matrix(), of which there were none.

Perfect duplicates were checked next, of which there were none.

Categorical variables were re-expressed using one-hot encoding via pandas .get\_dummies() function. Categorical variables with only binary decisions were re-expressed as 0 and 1 without dummy variables. The original columns for the dummy variables were dropped, as well as all columns unnecessary for the data analysis.

Pandas .corr() method and point biserial-r tests were used to determine correlation between Bandwidth\_GB\_Year and the remaining variables to get the final set of variables to be used for the model.

*4.  Provide a copy of the cleaned data set.*

File attached as “NCina D209 T2.csv”

***Part IV: Analysis***

*D.  Perform the data analysis and report on the results by doing the following:*

*1.  Split the data into training and test data sets and provide the file(s).*

Code attached as “NCina D209 T2.ipynb” in #Part D1

Files attached as “NCina D209 T2 train.csv” and ““NCina D209 T2 test.csv”

*2.  Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.*

Random forests were used to analyze the data. Random Search Cross Validation was used to determine the best hyperparameters to use for n\_estimators, max\_depth, and min\_samples\_leaf to achieve best model performance. The best performing hyperparameters were chosen using negative mean squared error as the performance metric, as the resulting model’s performance was validated using root mean squared error. No intermediate calculations were performed.

*3.  Provide the code used to perform the prediction analysis from part D2.*

Code attached as “NCina D209 T2.ipynb” in #Part D2

***Part V: Data Summary and Implications***

*E.  Summarize your data analysis by doing the following:*

*1.  Explain the accuracy and the mean squared error (MSE) of your prediction model.*

The model’s root mean squared error (RMSE) was ~242.05. The RMSE is the square root of the mean squared error (MSE) and represents the average distance of predictions from the real value. RMSE is easier to interpret than MSE as it is in the same units as the target variable. This means that, on average, the model predicts user gigabyte usage about 242.05gb more or less than the actual value. Considering the target variable ranges from 155gb to 7158gb, this is about 3.5% of the total range of the target variable away from the exact value and is therefore acceptably accurate.

An additional metric for evaluating random forests is R^2, of which the model achieves a score of ~0.988, which is exceptionally good.

*2.  Discuss the results and implications of your prediction analysis.*

The set of variables used in the analysis appears to be an accurate predictor of user gigabyte usage, implying there is a strong correlation between these variables and the amount of the service consumed by a user.

*3.  Discuss****one****limitation of your data analysis.*

Random forests are not good models for extrapolating data outside of the existing range *(Thompson, 2019).* If it were desired to predict customer bandwidth usage higher than 7158gb or below 155gb, random forests would perform much worse than other models that are good at extrapolation, such as linear regression.

*4.  Recommend a course of action for the real-world organizational situation from part A1 based on your results and implications discussed in part E2.*

If the company charges customers based on bandwidth usage, this model can be used to determine which services provide the most value to the company and what to push more advertising for. Additional research could be done into what types of customers are more likely to be purchasing those services.

***Part VI: Demonstration***

*F.  Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the programming environment.*

Panopto Video: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=5422b825-94c1-4069-8f34-b1b0000cce70>

*G.  Acknowledge web sources, using in-text citations and references, for segments of third-party code or data used to support the analysis. Be sure the web sources are reliable.*

Thompson, B. (2019, December 17). *A limitation of Random Forest Regression*. towardsdatascience. <https://towardsdatascience.com/a-limitation-of-random-forest-regression-db8ed7419e9f>

*H.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.*

No sources used.