Noah Cina

7/8/2024

D209 – Machine Learning

Performance Assessment Task 1

***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 classification methods:*

*•*k*-nearest neighbor (KNN)*

*•   Naive Bayes*

Given the current dataset, can customer churn for new customers be accurately predicted with a k-nearest neighbor classification 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 Churn and to produce a model that can predict if new customers will churn based on those variables with acceptable accuracy.

***Part II: Method Justification***

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

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

The classification method chosen is k-nearest neighbors, which determines the label of a datapoint by looking at the *k* closest existing datapoints’ labels in feature space and taking a majority vote. K-nearest neighbors makes binary decisions, which means that the outcome of this operation can only be a yes or no answer.

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

K-nearest neighbors assumes that a relationship exists between the datapoints in feature space and this relationship can be measured by distance metrics.

*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, standardization, pipelining, and assessing model validation statistics.

Numpy – standard data science library for python. Supports basic data structures used in data processing like arrays.

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.

***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 classification method from part A1.*

K-nearest neighbors requires 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 classification question from part A1 and classify*each*variable as numeric or categorical.*

The variables used for the classification model are as follows:

Categorical:

Techie

Phone

Multiple

OnlineBackup

DeviceProtection

StreamingTV

StreamingMovies

Gender\_Female

Gender\_Male

Contract\_Month-to-month

Contract\_One year

Contract\_Two Year

InternetService\_DSL

InternetService\_Fiber Optic

InternetService\_None

PaymentMethod\_Electronic Check

Numeric:

Tenure

MonthlyCharge

Bandwidth\_GB\_Year

*3.  Explain*each*of 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.

Chi-square contingency tests and point biserial-r tests were used to determine correlation between Churn and the remaining variables to get the final set of variables to be used for the model.

The data is not scaled at this step because scaling the numeric data handled by the modelling pipeline in part D2.

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

File attached as “NCina D209 T1.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 T1.ipynb” in #Part D1

Files attached as “NCina D209 T1 train.csv” and ““NCina D209 T1 test.csv”

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

K-Nearest Neighbors was used to analyze the data. Given the discrete nature of the hyperparameter “n\_neighbors”, grid search cross validation was used to determine the best k value for the model. The model was preprocessed using a pipeline to scale the continuous variables and increase model accuracy. Model performance was measured using a confusion matrix, scikit-learn’s classification report, and the area-under-curve score. No intermediate calculations were performed.

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

Code attached as “NCina D209 T1.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 area under the curve (AUC) of your classification model.*

The classification report is pictured below:

A screenshot of a computer screen

Description automatically generated

The model is decently accurate, with an accuracy of ~87%. Of all cases where a customer actually churned, the model predicted 68% of them (recall). For all predictions of churn the model made, 81% of them were correct (precision). For a variable with about a 75/25 ratio imbalance, the model is decent at correctly predicting a customer will not churn.

The AUC statistic of the model is ~0.93. AUC is a measure of the ratio between true and false positives. This means the model has an 93% chance of correctly predicting whether or not a customer is going to churn, which is 86% better than a random guess.

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

It appears that there is a correlation between churn and the selected variables used in the analysis, and that churn can be predicted decently accurately using a K-Nearest Neighbors classifier model. Given the AUC of 93%, the correlation can be predicted at about that rate for new customers.

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

The model is a little lacking in predicting whether a customer *will* churn (recall on the positive class), which can be expected from the 3 to 1 imbalance in the distribution of the Churn variable. A K-Nearest Neighbor model is not best suited for predicting reasonably rare events. It could be acceptable in this case given the large sample size and relatively forgiving imbalance, but a different model such as classification trees may be better suited for predicting customer churn.

*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.*

The model could be used to predict which customers are most likely to churn before they’ve done so, and action could be taken to prevent customer churn. Additionally, it is recommended to research the causes of customer churn and how to prevent it using the model predictions as a starting point.

***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 Link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=d042fcdd-041d-4bb5-97ec-b1a90185270e>

*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.*

No web sources used.

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

No additional sources used.