Patrick Callahan, Arjun Chimni, Andrew Schulz

CSCI 184: Machine Learning

4.15.22

Report: One-Hot Encoding

* The first step to the process was loading the data into a Jupyter Notebook and dealing with the missing values. Here, the approach we used was to simply delete the columns with Null values. Refer to the screenshots:
  + Table

    Description automatically generated
  + Graphical user interface, text, application, email

    Description automatically generated
  + Table

    Description automatically generated
* The next step to our process was to separate the data from input features and target features which was simple.
* After that, I split the data into only numerical and only categorical.
* Once the data was into two sets, the approach I used for the categorical data was to use the ‘get\_dummies’ function in Pandas so that each categorical column created a new dataframe that was one-hot encoded. Once this dataframe was made, I added it to a final categorical one-hot encoded data frame of roughly 159 columns.
* I appended this dataframe to the numerical dataframe. After this I had two sets of dataframes that could be used to build the LR models.
  + Table

    Description automatically generated
  + Table

    Description automatically generated
  + Table

    Description automatically generated
  + Table

    Description automatically generated
* Once I had the two sets (numerical data only and one-hot encoded categorical/numerical) of data, I was ready to build the simple Linear Regression models.
* I trained the data using sklearn library and observed the difference in Mean Absolute Error of the models.
* With the categorical data, the Mean Absolute Error decreases, thus showing the importance of one-hot encoding categorical data in a dataset.
  + Graphical user interface, text, application, email

    Description automatically generated