Building Predictive Models to Predict Occurrence of Stroke

Shivam Patel

Summary:
The purpose of this project was to build predictive models that predict the likelihood a person has a stroke based on various given variables and find the more optimal model by referencing the coefficient of determination (R2) and root-mean-square error (RMSE) that the individual model produced. The predictive models were built using Python and model building libraries such as Sklearn. A testing and training datasets were provided where the training dataset was used to train the models while the testing dataset was used to test the models. The variables in the datasets included: ID, Gender, Age, Hypertension, Heart Disease, Ever Married, Work Type, Residence Type, Average Glucose Level, BMI, Smoking Status, and Stroke. Both datasets were cleaned using various wrangling techniques which handled null values and outliers. The first predictive model that was built was linear regression where various models differing in variable usage were run. The highest R2 and RMSE values that the linear model was able to produce were 0.0702 and 0.1861, respectively. This linear model used both the categorical variables and numeric variables. The second model build was a classification tree where the highest R2 and RMSE values that were produced were -0.0403 and 0.1968, respectively. This classification tree had a depth of 3. Finally, the last model that was built used a k-nearest neighbor algorithm which produced a R2 and RMSE of -0.0129 and 0.1942, respectively. Overall, the results were not as expected. Specifically, a base line R2 and RMSE of 0.872 and 0.2060 were provided based on a simple linear model, however, the predictive models in this project did not reach this baseline, though the linear regression model came close.

Data:
The data used for this project was provided between 2 datasets, a testing dataset and a training dataset. The training dataset was used to train all the predictive models while the testing dataset was used to test the models. The data had various variables such as

- ID, a numerical variable to keep track of each individual patient
- Gender, a categorical variable for the gender of the patient which had three options: ‘Male’, ‘Female’, or ‘Other’
- Hypertension, a numerical variable for whether the patient had high blood pressure or not which was 0 indicating a no and 1 indicating a yes
- Heart Disease, a numerical variable for whether the patient had heart disease or not which was 0 indicating a no and 1 indicating a yes
- Ever Married, a categorical variable for whether the patient was married or not where Yes indicated they were and No indicated they were not
- Work Type, a categorical variable for the type of job the patient had which was broken into 5 categories: ‘Govt_job’ which indicated federal employment, ‘Self-employed’ which indicated self-employment, ‘Private’ which indicated private sector employment, ‘children’ which indicated homemaker, and ‘Never_worked’ which indicated no employment
- Residence Type, a categorical variable for the area the patient resides in which was ‘Urban’ indicating an urban area or ‘Rural’ indicating a rural area

- Average Glucose Level, a numerical variable which a measure of the patient's average glucose level
- BMI, a numeric variable which was a measure of the patient's BMI index according to the BMI scale
- Smoking Status, a categorical variable which indicates the patient's smoking status in which 'formerly smoked' indicates that the patient smoked before but not currently, 'never smoked' indicates that the patient has never smoked , 'Unknown' indicates that the patient's smoking status is not known, and 'Smokes' which indicates that the patient currently smokes.
- Stroke, a numeric variable which indicates whether the patient had a stroke where 0 indicates that they did not and 1 indicates that they did.

Both datasets were read in and checked for null values. Null values for BMI were found in the datasets and those rows were dropped. The Bi-categorical variables such as Gender, Ever Married, and Residence Type were all converted into numeric by replacing the choices with 0 or 1. Box and Whisker plots were plotted for numeric variables such as Average Glucose Level and BMI to see if there were any outliers. The outliers were then windsorized and the new values were stored in separate columns named 'avg_glucose_level_wind' and 'bmi_wind'. After cleaning the data, a testing and training dataframe was made where the variable ID was dropped as it has no relation with the data besides keeping track of each person.

Results:

The first predictive model made was a Linear Regression model. A function to calculate the RMSE and a function to calculate the R2 values were coded as well as a function that performed linear regression with no expansion. First, the model was run with only numeric variables such as avg_glucose_level_wind',' bmi_wind', 'age', 'hypertension',' heart_disease', 'ever_married', 'Residence_type', and 'gender'. The R2 and RMSE values from this model were 0.0640 and 0.1867, respectively. The model was then run with only categorical variables such as 'work_type' and 'smoking_status' and the R2 and RMSE values generated from this model were 0.0081 and 0.1922, respectively. Finally, the model was run using both the numeric and categorical variables which gave an R2 and RMSE values of 0.0702 and 0.1861, respectively. Since the results did not reach the base line R2 and RMSE values provided from a simple linear model 0.0872 and 0.2060, respectively, another function that performed linear regression with expansion was coded. The different combination of variables was retested with this new model. The numeric variables with expansion produced an R2 and RMSE value of 0.0558 and 0.1875, respectively. The categorical variables with expansion produced an R2 and RMSE value of 0.0060 and 0.1924, respectively. The combination of these variables expanded produced R2 and RMSE values of 0.0532 and 0.1878, respectively. These expanded models performed worse than the simple linear models, therefore, a new strategy of selecting variables based on correlation with stroke was performed. The correlation for each variable with stroke was calculated (Figure A). The most correlated variables, 'age', 'hypertension', 'heart_disease', and 'avg_glucose_level_wind', had a correlation value of 0.2311, 0.1458, 0.1430, and 0.1241, respectively. The linear regression with no expansion was run with these variables and R2 and RMSE values produced were 0.0624 and 0.1868, respectively. Even with using the most correlated variables, the base line R2 and RMSE values were not reached.

```
Correlation of each Categorical variable with 'stroke':
id                                  0.000157
gender                             -0.009877
age                                 0.231056
hypertension                        0.145836
heart_disease                       0.142955
ever_married                        0.097515
Residence_type                     -0.002063
avg_glucose_level                   0.140527
bmi                                 0.039494
stroke                              1.000000
avg_glucose_level_wind              0.124055
bmi_wind                            0.046783
work_type_Never_worked             -0.013212
work_type_Private                   0.011569
work_type_Self-employed             0.058798
work_type_children                 -0.080439
smoking_status_formerly smoked      0.051100
smoking_status_never smoked         0.007368
smoking_status_smokes               0.034242
Name: stroke, dtype: float64
```

Figure A: Correlation of each variable with the variable ‘stroke.

The second predictive model made was a Classification Tree. The RMSE and R2 functions were edited to be able to calculate their respective values with the tree model. The tree model was run with various depths to find the most optimal tree model. At a depth of 3 the R2 and RMSE values were -0.0403 and 0.1968, respectively. At a depth of 5, the R2 and RMSE values were -0.1498 and 0.2069, respectively. At a depth of 10, the R2 and RMSE values were -0.2319 and 0.2142, respectively. Finally, at a depth of 20, the R2 and RMSE values were -0.9163 and 0.2574, respectively. Given the fact the model produced worse R2 and RMSE values as the depth of the trees increased, a model with a depth greater than 20 was not attempted. The best tree model was with a depth of 3 and its R2 and RMSE values did not reach the baseline values and furthermore, the R2 value was negative which indicated that the tree model was not a great predictive model. The third and last predictive model was made by using k-Nearest Neighbors (knn) library from sklearn. First the training data and testing data was normalized using a max-min normalization function. A function that performs the knn algorithm was created and the R2 and RMSE functions are updated in order to be able to calculate their respective values with the knn model. Using a for loop, the knn model was run with up to k = 70, to find the optimal number of nearest neighbors. The optimal number was the model with the k value that produced the lowest RMSE

value. The optimal k value was found to be 9. The knn algorithm was rerun with k = 9 and the R2 and RMSE values that were produced were -0.0129 and 0.1942, respectively. The knn model even with the optimal number of neighbors did not reach the baseline R2 and RMSE values and furthermore, the R2 was negative indicating that the knn model was not a great predictive model. A table of each of the 3 models' R2 and RMSE values were created (Figure B).

	R2	RMSE
Linear Regression	0.070202	0.186069
Classification Tree	-0.040297	0.196815
K-Nearest Neighbor	-0.012921	0.194208

Figure B: Table of each predictive models' best R2 and RMSE values.

As can be seen from the table in figure B, the most successful predictive model was the linear regression model which had the highest R2 value (0.0702) and the lowest RMSE (0.1861). This model is followed by the k-nearest neighbor model and lastly, the classification tree. Therefore, the best predictive model made was the linear regression model even though it still did reach the baseline R2 and RMSE values provided.

Conclusion:
The project aimed to make and test various predictive models for stroke occurrence based on a variety of variables and using different machine learning techniques. Three models were built and used: linear regression, classification tree, and k-nearest neighbors. The evaluation of each model was based on the highest R2 and lowest RMSE values. The optimal linear regression model was found by testing different degrees of expansion and using various combinations of variables. The most optimal linear regression model was made using a combination of the numeric and categorical variables with no expansion. The optimal classification tree model was found by testing various depths. The most optimal tree model was made with a depth of 3. Finally, the optimal k-nearest neighbor model was found by running the model with various k values. The most optimal k-nearest neighbor model was made using 9 neighbors. The best R2 and RMSE values were produced from the linear regression model with R2 being 0.0702 and the RMSE being 0.1861. However, none of the models reached the baseline R2 and RMSE values reached by the simple linear model. Overall, the models fell short of expectations as the baseline values were not reached. The tree and knn models exhibited poorer predicative capabilities with a negative R2 value. Despite the models' inability to surpass the baseline values, the projected adapted a systematic approach. It encompassed data cleaning, exploration, and feature selection based on correlation and rigorous testing of different methods. Optimization for each model was

attempted, however, the models may have been limited by the data itself. Firstly, strokes difficult to predict “because the data related to the disease are multi-modal.” (Liu et al., 2020). Therefore, using variables such as “smoking” and “average glucose level” to predict the likelihood of a stroke occurring and getting accurate results is difficult. The complexity of predicting stroke occurrence with accuracy may require more robust variables such as a patient’s medical history or heredity. More robust predictive algorithms may also be required.

Work Cited

Liu, Y., Yin, B., & Yang, C. (2020). The Probability of Ischaemic Stroke Prediction with a Multi-Neural-Network Model. *Sensors*, *20*(17), 4995–4995. https://doi.org/10.3390/s20174995