STROKE PREDICTION

Varun Reddy

BACKGROUND

Stroke is the second leading cause of death in the world, accounting for nearly 11% of all deaths, according the World Health Organization (WHO).

A stroke occurs when blood supply is cut or reduced to a portion of the brain, thereby reducing oxygen to that particular area. Within minutes of losing blood supply, brain cells begin to die.

PROBLEM STATEMENT

We were tasked to predict whether a patient will have a stroke episode based on 11 various features such as gender, age, hypertension, and bmi. These findings will be used to flag potential high-risk patients for physicians to further investigate.

We maximized for the Recall metric to make sure that we do not falsely identify patients as a non-stroke risk. False positives are not as much of a concern because of the intended use of the model to identify potential high-risk patients.

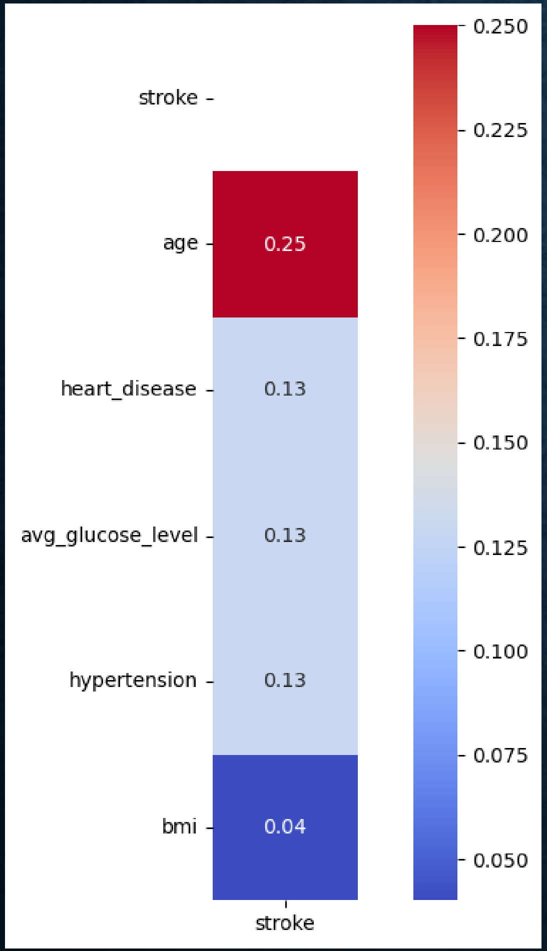
STROKE DATA

Data was taken from Kaggle Stroke Dataset.

Data Details:

- 11 features – gender, age, hypertension, heart_disease, marriage, work type, residence, avg glucose level, bmi, smoking status, stoke episode
- 5,110 patient samples
- 4,861 non-stroke patients, 249 stroke patients
- 201 missing BMI values

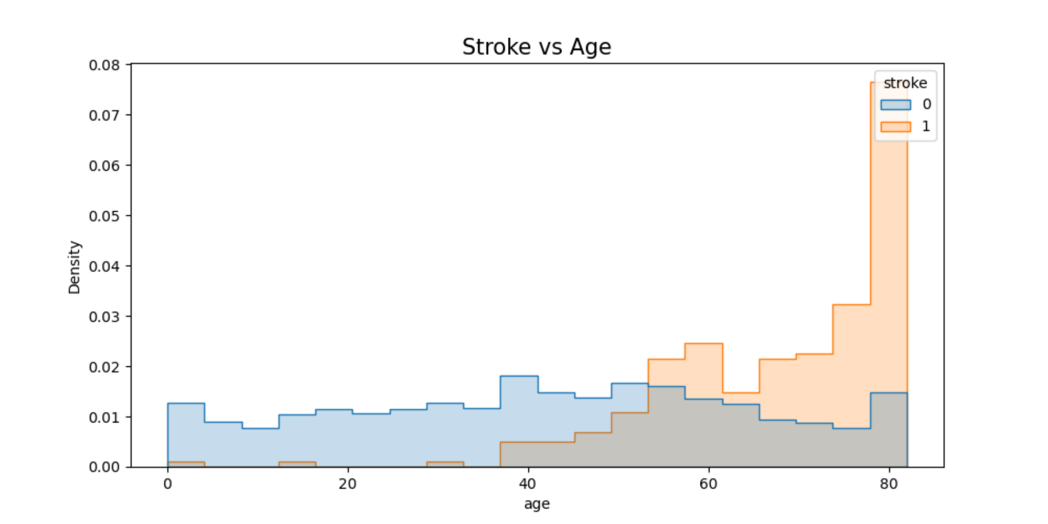
DATA CORRELATION

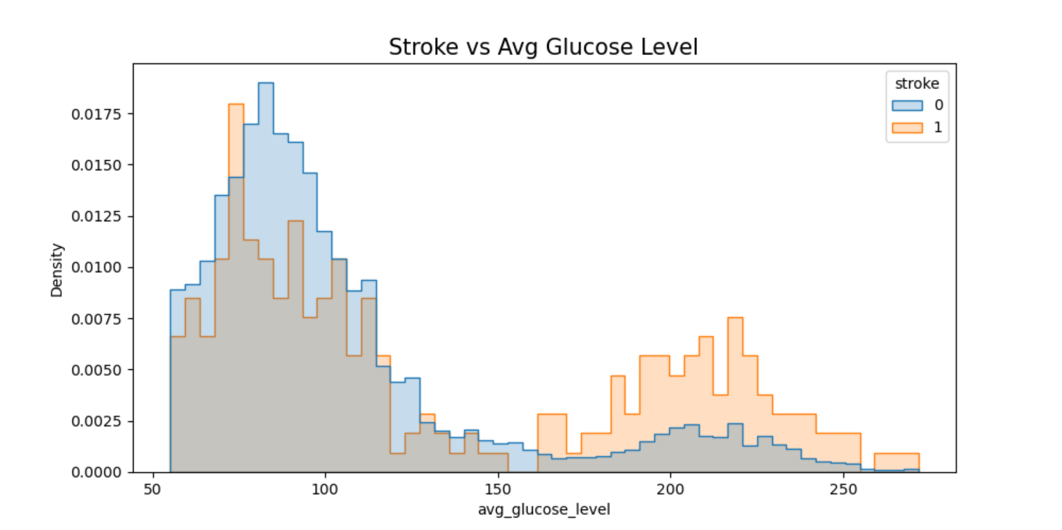


The highest correlated features are age, heart disease, and avg. glucose levels

DATA EXPLORATION

Clear correlation as age increase the risk of stroke also increases.





Above an average glucose level of 150 the risk of stroke increases.

IMBALANCED CLASSES & FEATURE ENGINEERING

With 4,861 non-stroke patients and 249 stroke patients, the classes were severely imbalanced with nearly 95% of patients belonging to the majority class and 5% of patients belonging to the minority class. Synthetic Minority Oversampling Technique (SMOTE) was used to balance the training set.

Polynomial Features with n=3 was employed to explore more relationships between features and target.

MODEL PERFORMANCE

Model	Accuracy	Recall	Precision	F1
Naïve Bayes	**.72**	**.68**	**.11**	**.16**
Logistic Regression	.80	.39	.10	.16
KNN	.84	.26	.09	.13
Random Forest	.94	.10	.27	.14
Ada Boost	.94	.06	.17	.09
Stacked	.95	0	0	0
NULL	.95	0	0	0

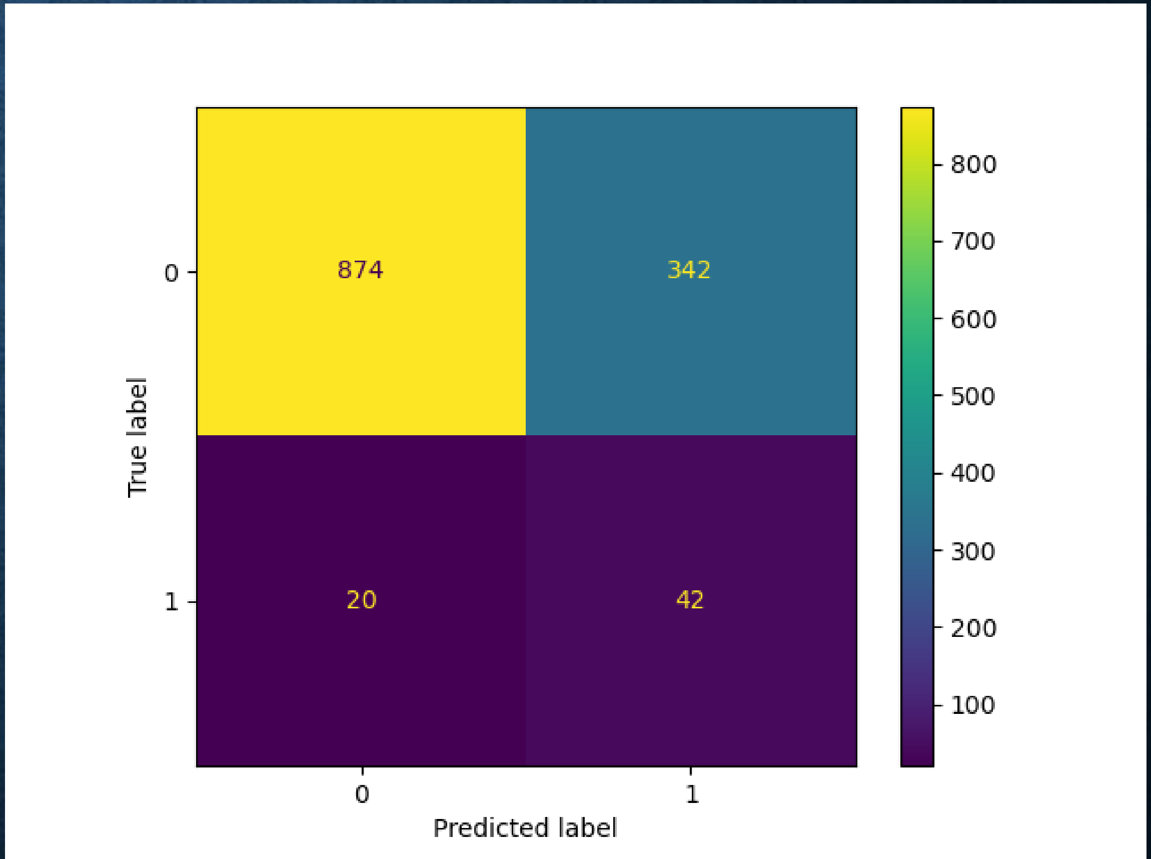
Polynomial Features were used in models to create 969 total features.

Models optimized for Recall and reduce false negatives.

OPTIMAL MODEL – NAÏVE BAYES

Naïve Bayes with Polynomial Features (n=3) was chosen due to highest Recall metric.

Recall – 68%
Precision – 11%
Accuracy – 72%



CONCLUSION AND NEXT STEPS

Although Naïve Bayes with Polynomial Features has a lower Accuracy score than the Null Model it has a significantly higher Recall score. This can be helpful in screening for high risk patients for further investigation while reducing false negatives.

Future steps can be to further refine the models to increase for both Recall and Accuracy. The distributions of highly correlated features especially average glucose level can be examined and adjusted to increase model performance for Recall.