

IOD Mini Project 2

Regina Soh



Content

- Housing price prediction
 - Linear Regression
 - Lasso and Ridge Regression
- Stroke prediction
 - Logistic Regression
 - Support Vector Machine
 - K-nearest Neighbors

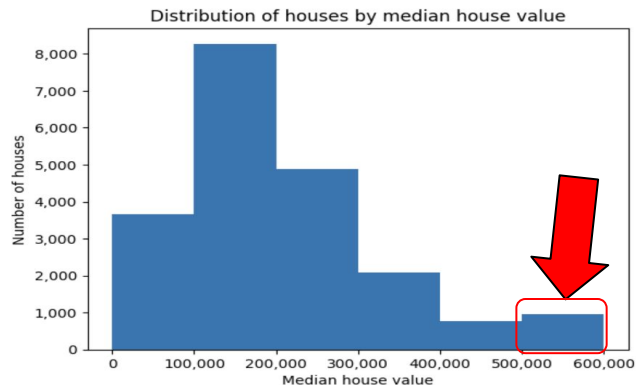
Housing Price Prediction



California Housing Prices Dataset

- Information on houses found in California based on the 1990 census data
 - Location: longitude, latitude, ocean_proximity
 - Property information: total_rooms, total_bedrooms, housing_median_age, median_house_value
 - Demographics: population, households
- Data source
 - <https://www.kaggle.com/datasets/camnugent/california-housing-prices>

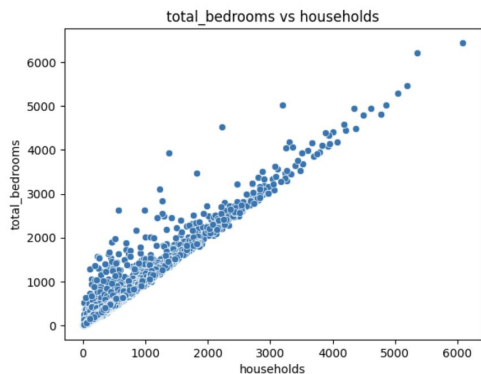
Data Cleaning



1) Remove data with house price at 500,001

965 houses with the same house value of 500,001

Very unlikely for so many houses to have the same median price, especially when it is a high price.

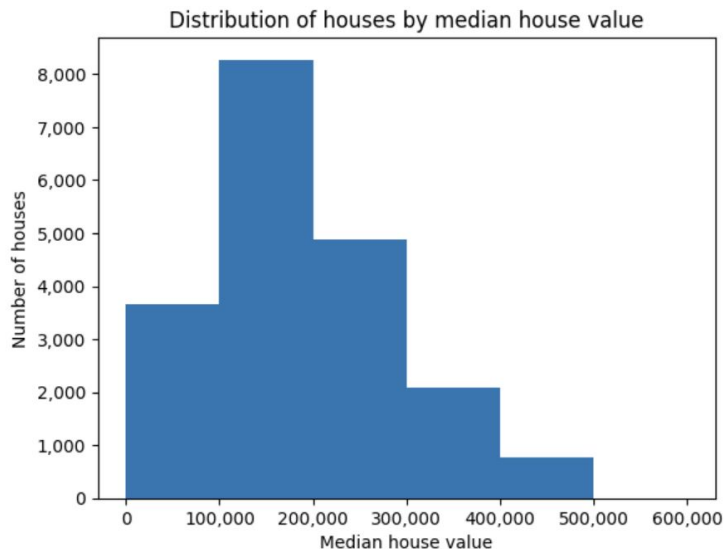


2) Estimate missing total_bedrooms data using linear regression with households

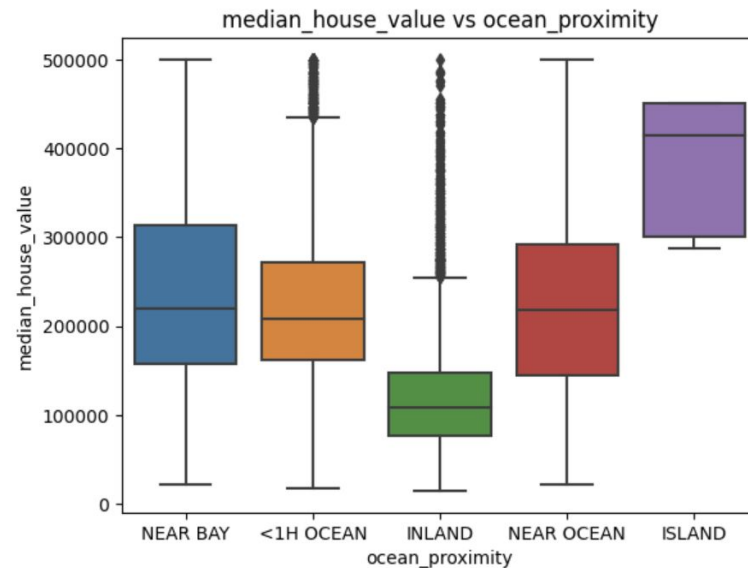
Household has a very strong linear relationship with total_bedrooms

Training result accuracy at 0.96

Exploratory Data Analysis

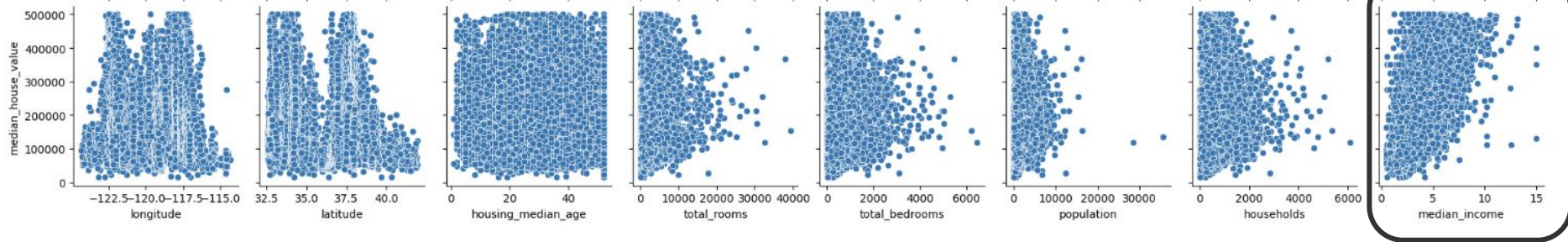


A large proportion of houses have median house value between 100k - 200k



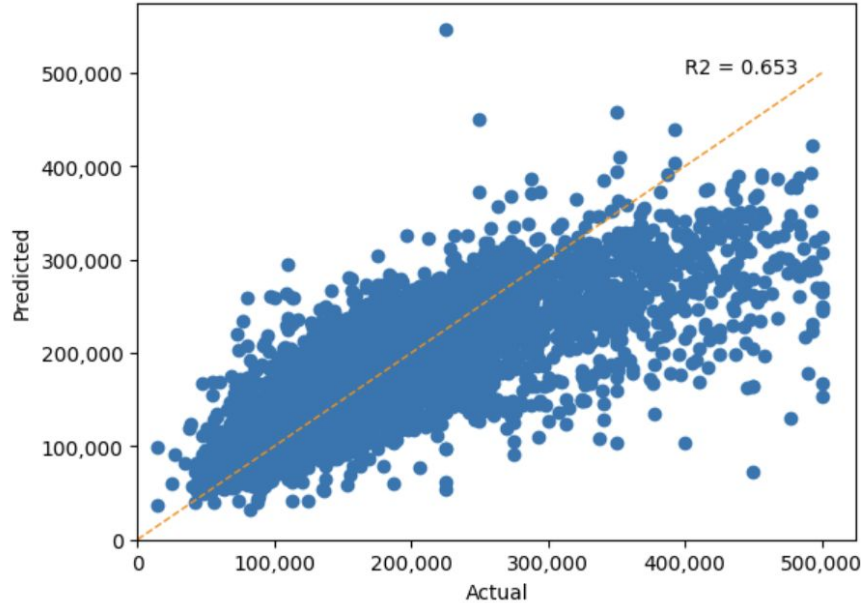
Houses on the island are more expensive while those in the inland are cheaper

Exploratory Data Analysis



Housing price has some linear relationship with median_income

Price Prediction



**Linear Regression model
could not predict higher
value houses well**

Training accuracy score: 0.641

Test accuracy score: 0.653

Comparing different models

TRAINING SCORES	r2	mse
lr	0.641114	0.101443
lasso	0.641114	0.101443
ridge	0.641114	0.101443

TEST SCORES	r2	mse
lr	0.652886	0.101138
lasso	0.652887	0.101138
ridge	0.652886	0.101138

All three models have the same predictive power

Possible Improvements: Get data on housing amenities (pool, fitness corner etc)

Stroke Prediction



Stroke Prediction Dataset

- Patients information
 - Basic information: id, gender, age
 - Health-related: stroke, hypertension, heart_disease, avg_glucose_level, bmi
 - Lifestyle: smoking_status, ever_married, work_type, Residence_type
- Data source
 - <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

Data Cleaning

1) Remove patients without BMI data

4% of total patients

#	Column	Non-Null Count	Dtype
0	id	5110 non-null	int64
1	gender	5110 non-null	object
2	age	5110 non-null	float64
3	hypertension	5110 non-null	int64
4	heart_disease	5110 non-null	int64
5	ever_married	5110 non-null	object
6	work_type	5110 non-null	object
7	Residence_type	5110 non-null	object
8	avg_glucose_level	5110 non-null	float64
9	bmi	4909 non-null	float64
10	smoking_status	5110 non-null	object
11	stroke	5110 non-null	int64

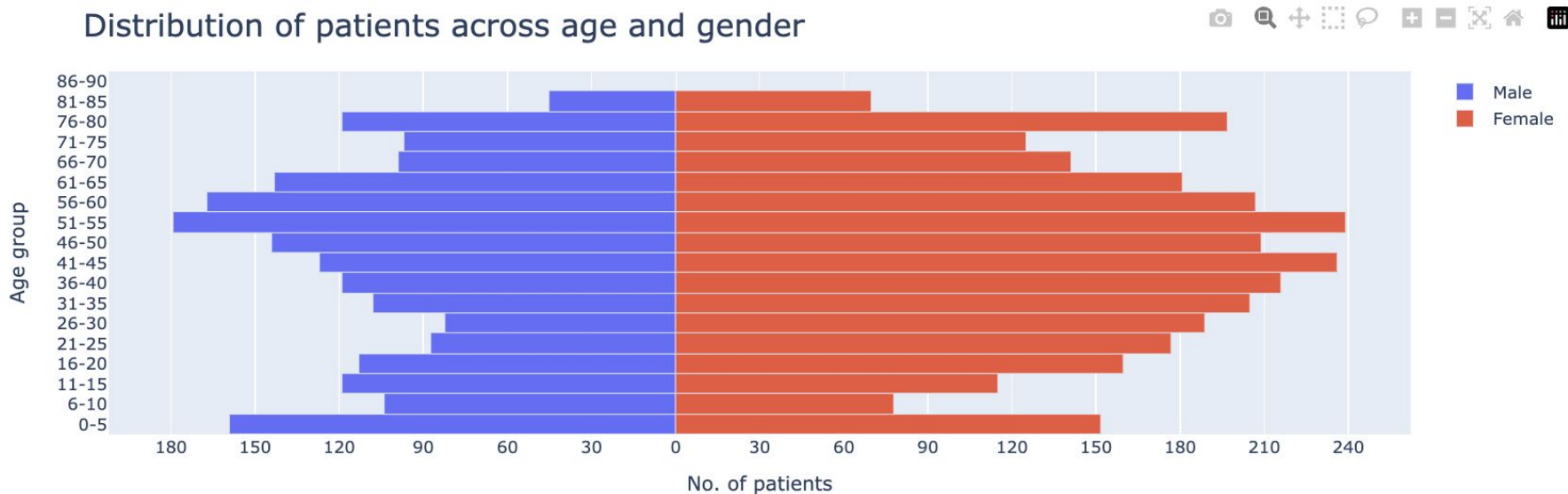
2) Remove patient with gender 'Other'

Only one patient with 'Other' gender

```
gender
Female    2897
Male      2011
Other         1
Name: count, dtype: int64
```

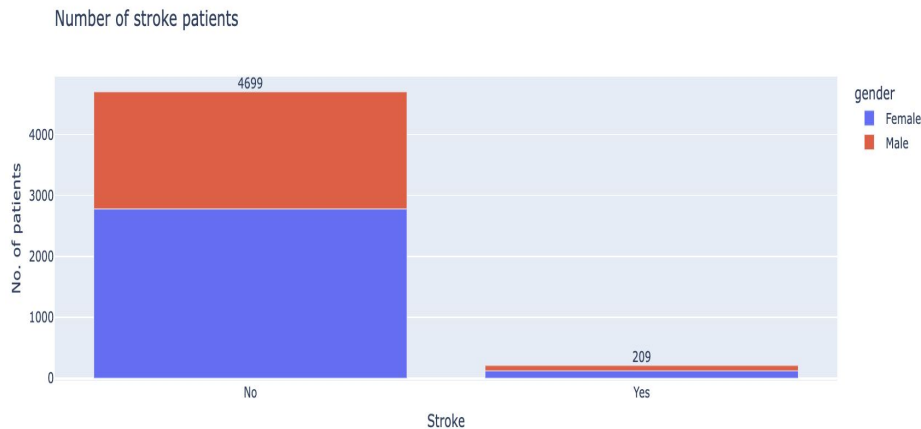
Exploratory Data Analysis

Distribution of patients across age and gender

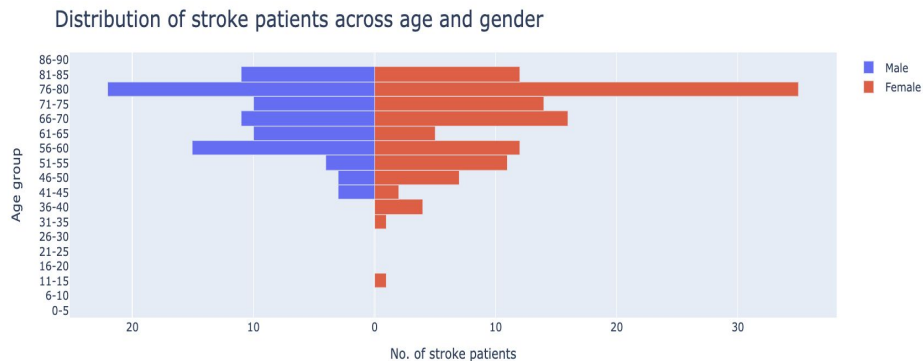


Distribution of patients across age group look similar for both genders

Exploratory Data Analysis



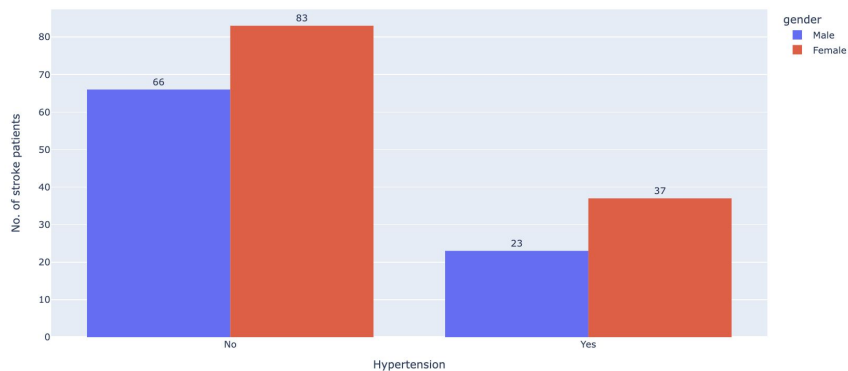
Very small number of stroke patients in dataset (4%)



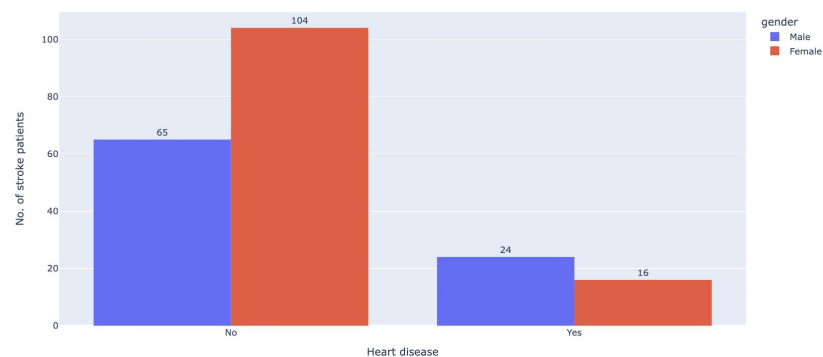
A large percentage of stroke patients are between the age 76-80 for both genders

Exploratory Data Analysis

Stroke patients with hypertension



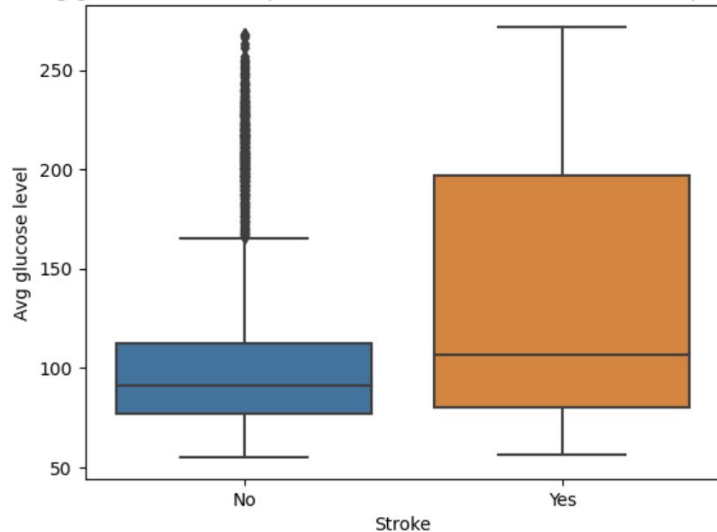
Stroke patients with heart disease



Majority of stroke patients do not have hypertension or heart disease

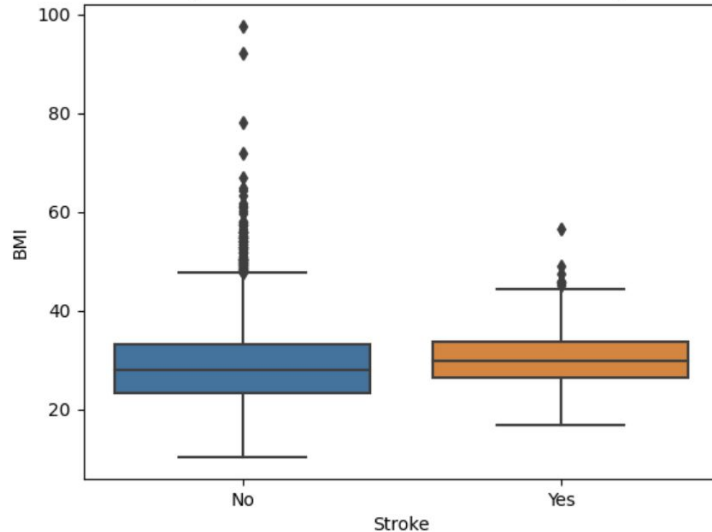
Exploratory Data Analysis

Avg glucose level comparison between non-stroke and stroke patients



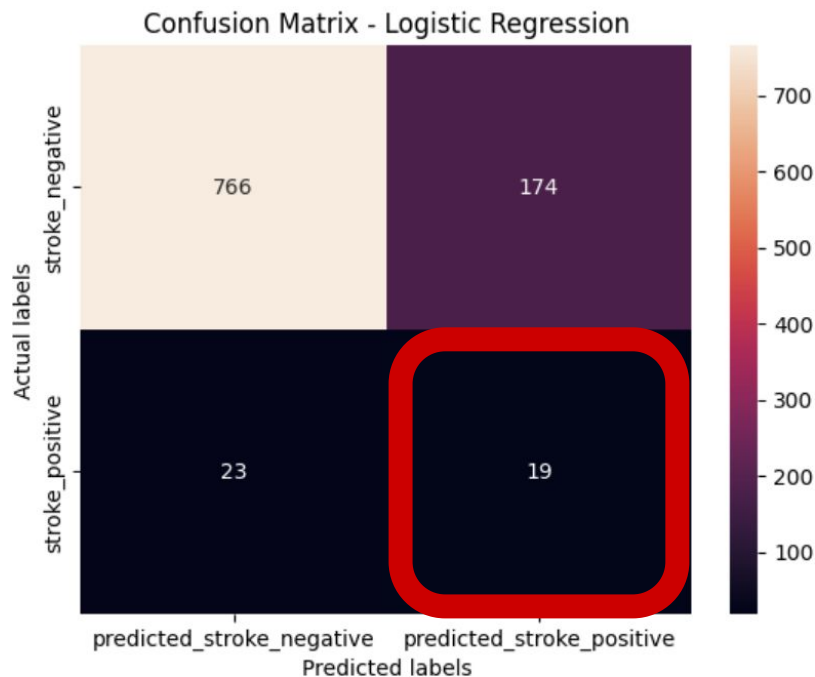
Higher avg glucose levels are more common in stroke patients

BMI comparison between non-stroke and stroke patients



Lower BMIs are more common in non-stroke patients

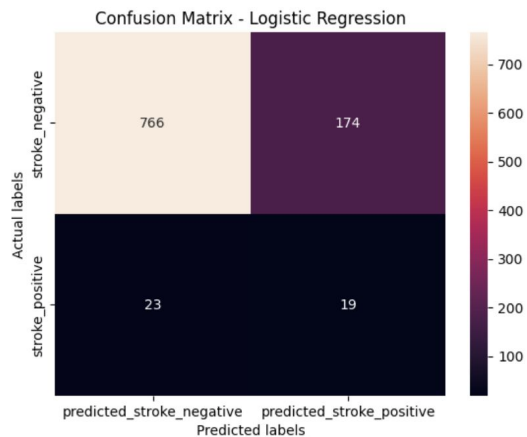
Stroke prediction



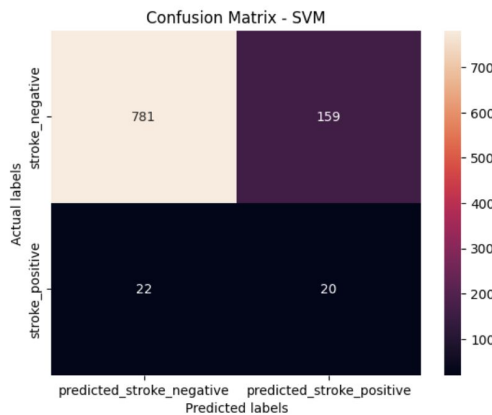
**Logistic Regression model
could not predict stroke
patients well**

Training F1 score: 0.843
Test F1 score: 0.162

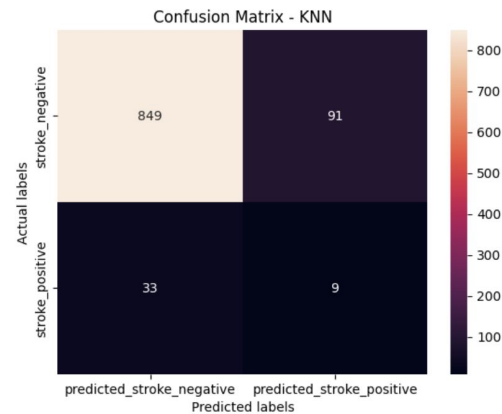
Comparing different models



Training F1 score: 0.843
Test F1 score: 0.162



Training F1 score: 0.861
Test F1 score: 0.181



Training F1 score: 1.000
Test F1 score: 0.127

All three models are poor at predicting stroke patients

How to improve models

Try different methods of handling imbalanced data

Models are likely overfitting training data

- Training score high but low test score
- Current model uses oversampling method to tackle imbalanced data

Get more stroke patients data

Real data is better than using algorithms to fix class imbalanced

- Model will be more trustworthy

End