IOD_mini_project2

Ng Jing Kang

IOD_mini_project2

Problem statements and datasets for this project:

Problem Statement	***Dataset***	***Description of data***	***Remarks***
Build a regression model to predict "median_house_value". Explore what we used in class.	https://www.kaggle.com/datasets/camnugent/california-housing-prices	This dataset is used to predict "median_house_value". Median house prices for California districts derived from the 1990 census.	-
Build a classification model to predict stroke (= 1 if a person had a stroke else 0). Explore what we used in class.	https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset	This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status.	-
Build a clustering model to segment customers. Explore what we used in class.	https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis	This dataset is used to cluster and analyze a company’s ideal customers.	WIP

IOD_mini_project2: Regression

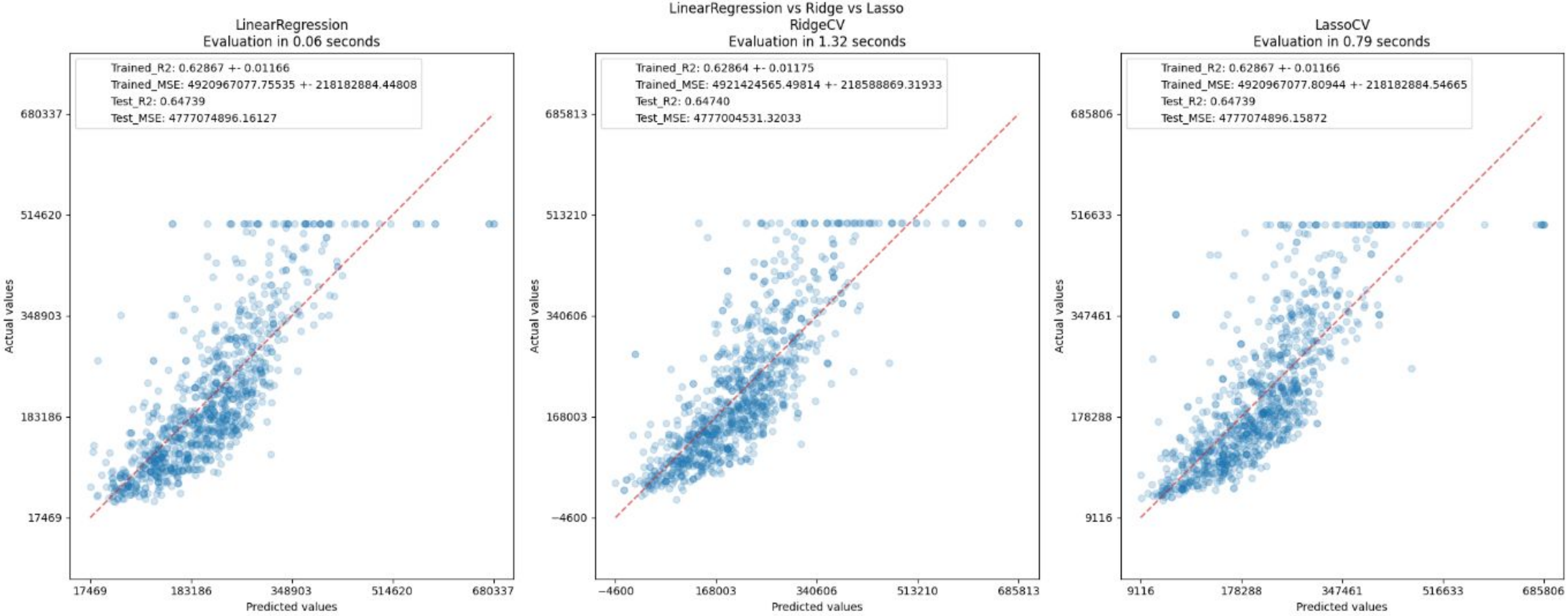
Dataset: https://www.kaggle.com/datasets/camnugent/california-housing-prices

Profiling report: housing_profiling.html

Columns	***Description***
median_house_value	Target for prediction
housing_median_age, total_rooms, total_bedrooms, population, households, median_income, ocean_proximity	Predictor columns
longitude, latitude	Not useful for prediction

IOD_mini_project2: Regression

- Regression model comparison



LinearRegression vs Ridge vs Lasso

LinearRegression
Evaluation in 0.06 seconds

```
Trained_R2: 0.62867 +- 0.01166
Trained_MSE: 4920967077.75535 +- 218182884.44808
Test_R2: 0.64739
Test_MSE: 4777074896.16127
```

RidgeCV
Evaluation in 1.32 seconds

```
Trained_R2: 0.62864 +- 0.01175
Trained_MSE: 4921424565.49814 +- 218588869.31933
Test_R2: 0.64740
Test_MSE: 4777004531.32033
```

LassoCV
Evaluation in 0.79 seconds

```
Trained_R2: 0.62867 +- 0.01166
Trained_MSE: 4920967077.80944 +- 218182884.54665
Test_R2: 0.64739
Test_MSE: 4777074896.15872
```

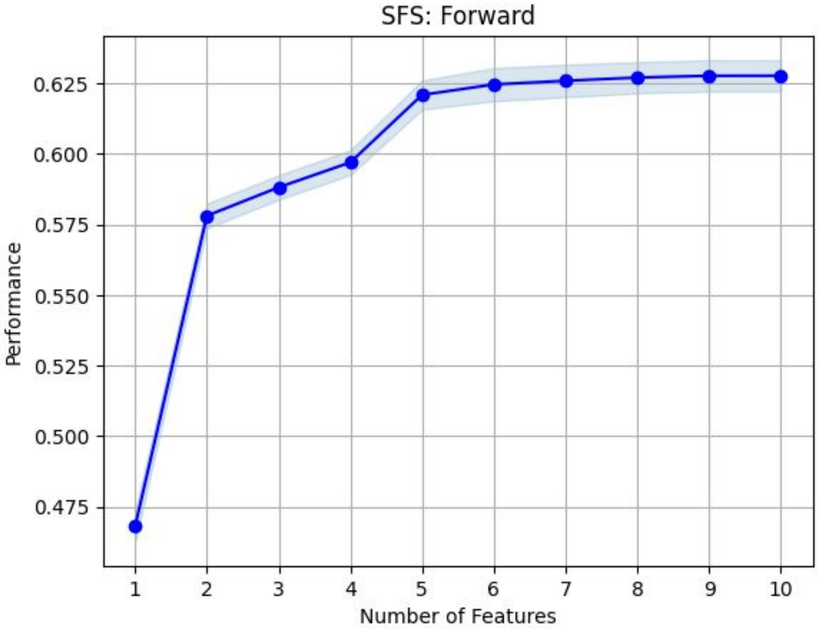
IOD_mini_project2: Regression

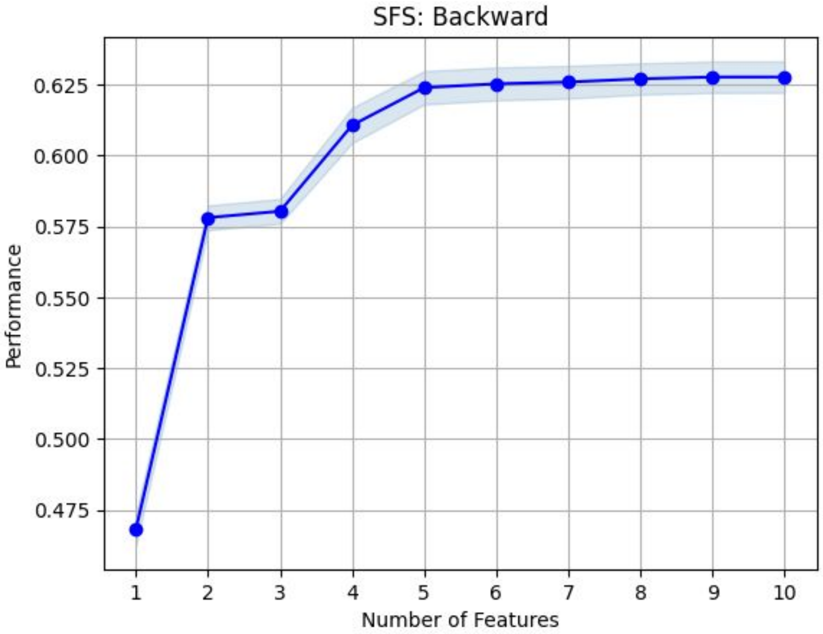
- Cross-validation estimator (RidgeCV, LassoCV)
  - An estimator that has built-in cross-validation capabilities to automatically select the best hyper-parameters
  - Roughly equivalent to GridSearchCV
  - They can take advantage of warm-starting by reusing precomputed results in the previous steps of the cross-validation process
  - Generally leads to speed improvements

Regression Models	***Description***
Linear Regression	OLS Linear Regression
RidgeCV	Ridge regression with built-in cross-validation. CV: Leave-One-Out Cross-Validation.
LassoCV	Lasso regression with built-in cross-validation. CV: 5 (default).

IOD_mini_project2: Regression

- Feature selection: Sequential feature selector


SFS: Forward


SFS: Backward

IOD_mini_project2: Classification

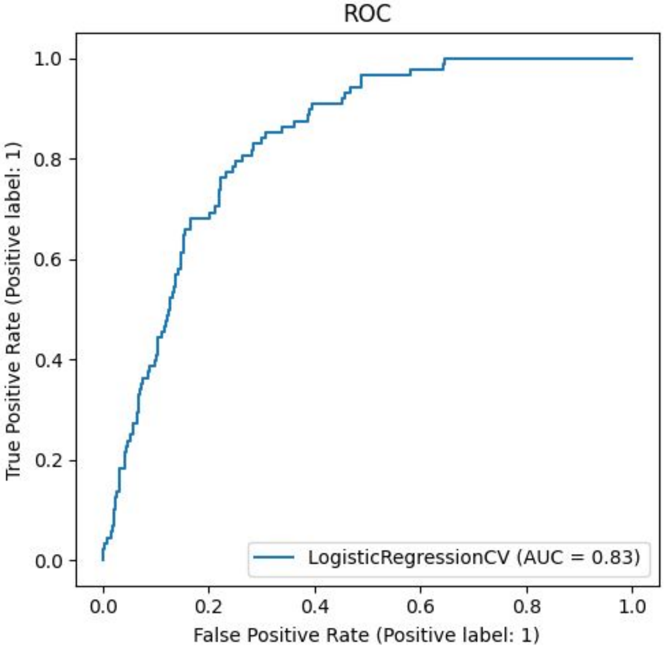
Dataset: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

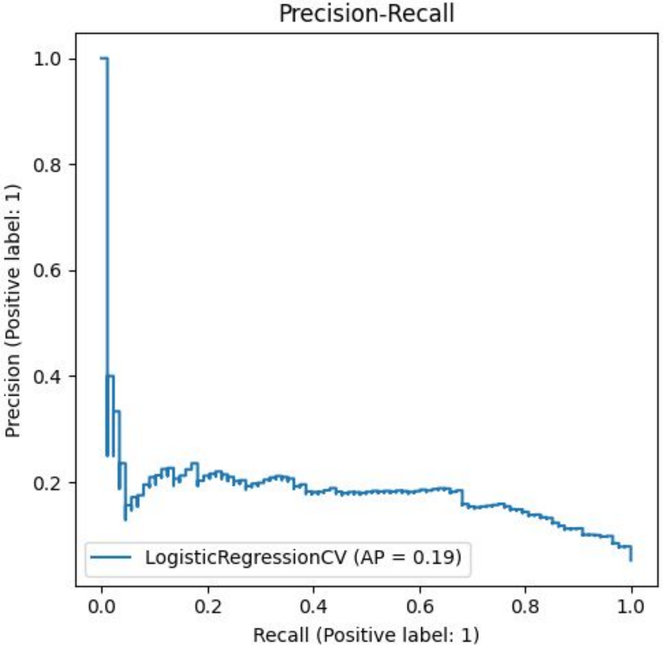
Profiling report: healthcare_profiling.html

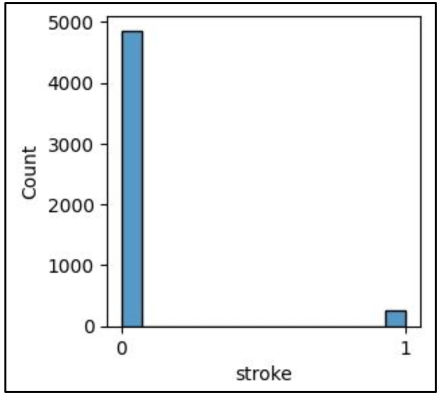
Columns	***Description***
stroke	Target for prediction
gender, age, hypertension, heart_disease, work_type, Residence_type, avg_glucose_level, bmi, smoking_status	Predictor columns
id, ever_married	Not useful for prediction

IOD_mini_project2: Classification

- Classification model: LogisticRegressionCV
  - No sampling strategy

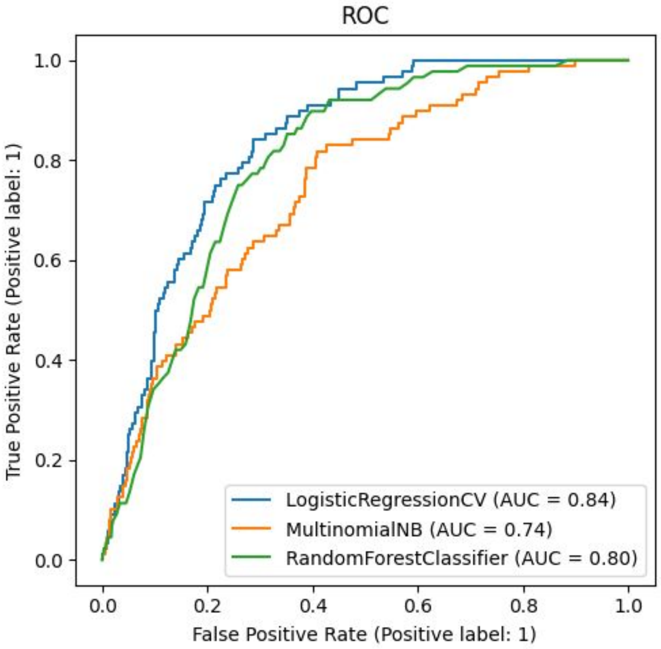


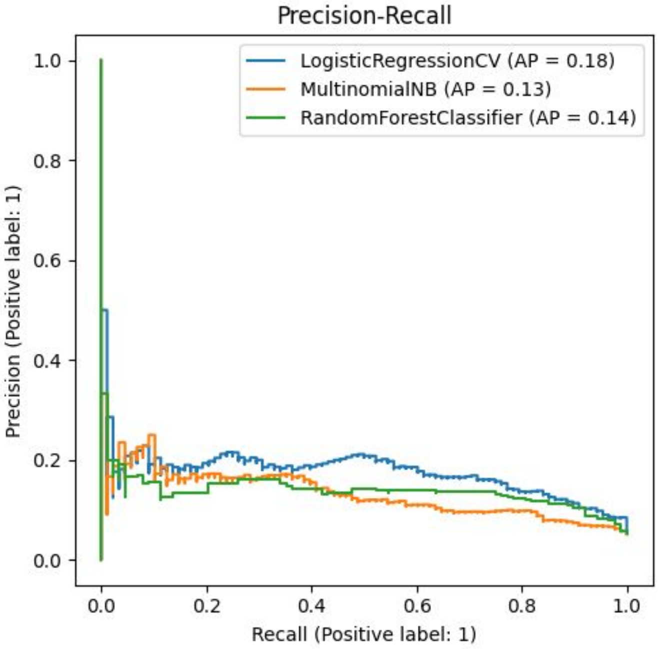


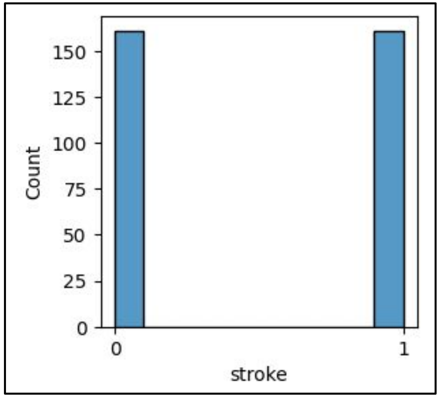


IOD_mini_project2: Classification

- Classification model comparison
    - Comparison of models’ performance using ROC / Precision-Recall curve
    - The logistic regression model has a highest roc-auc and precision recall ap and therefore will be concluded as the better classifier to use.

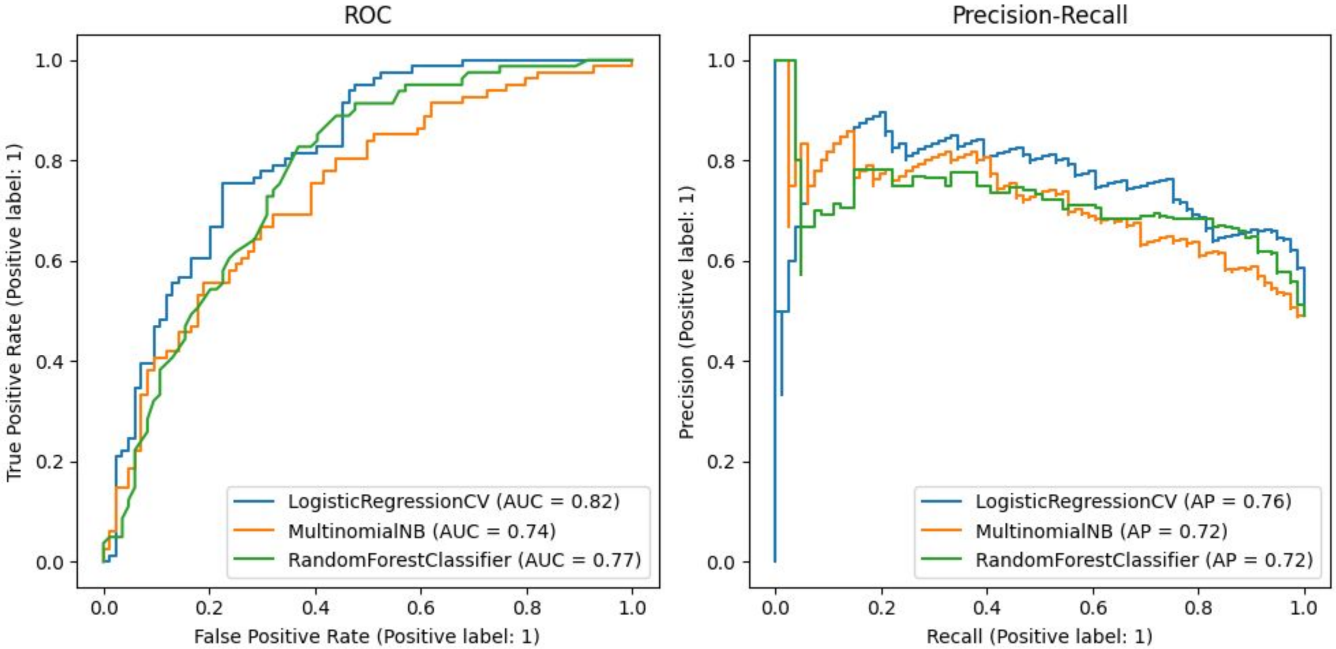


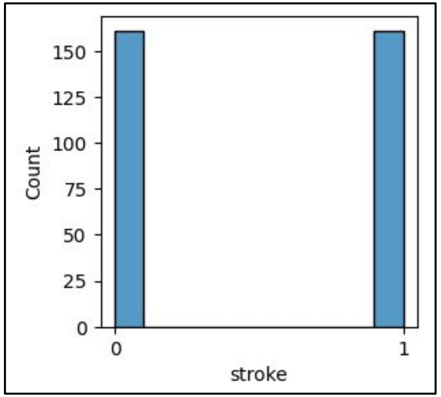




IOD_mini_project2: Classification - Wrong way to sample

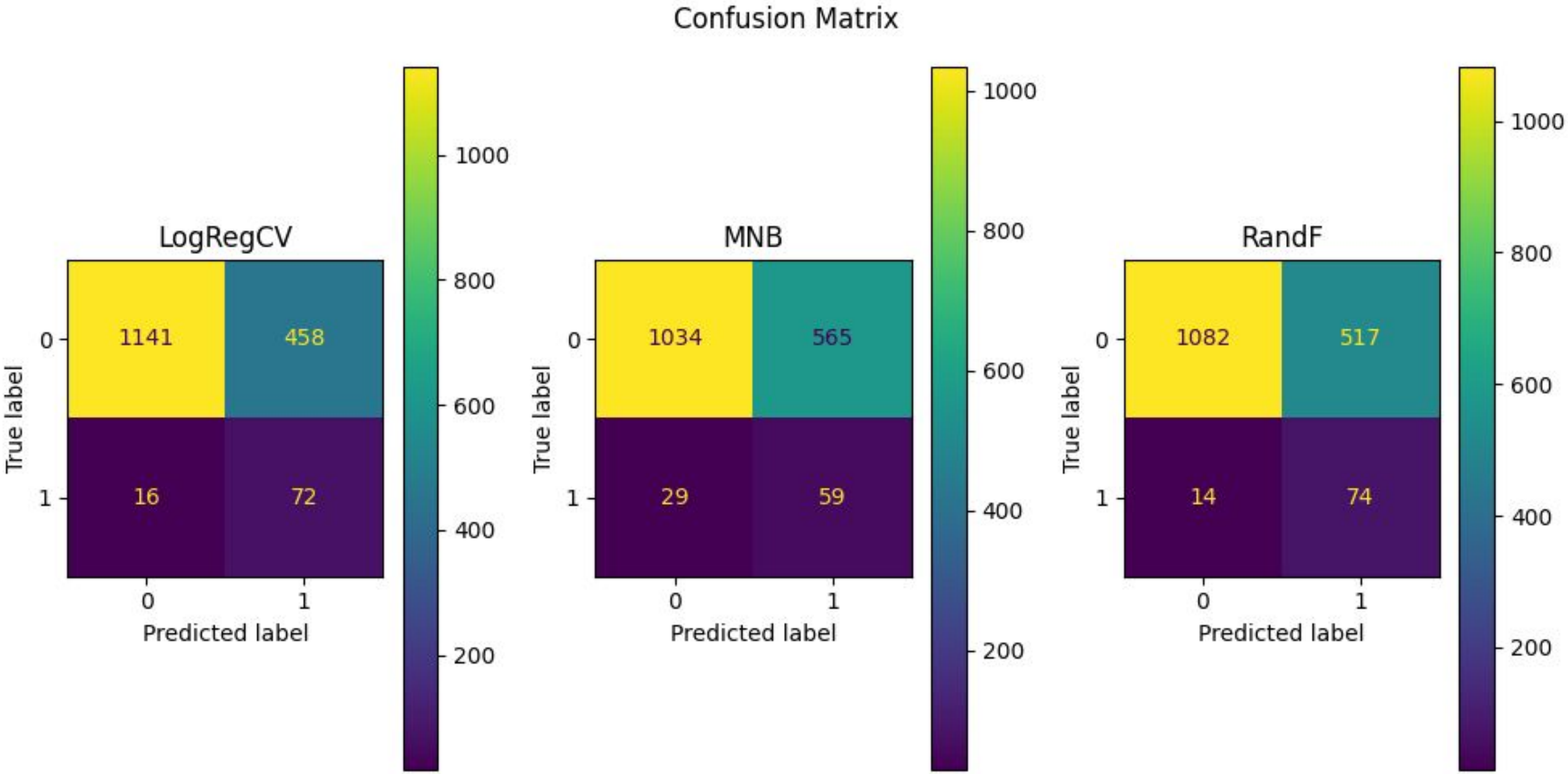
- Classification model comparison
  - Transform X,y with random under sampling
  - Split into X_train, y_train, X_test, y_test
  - Predict with X_test





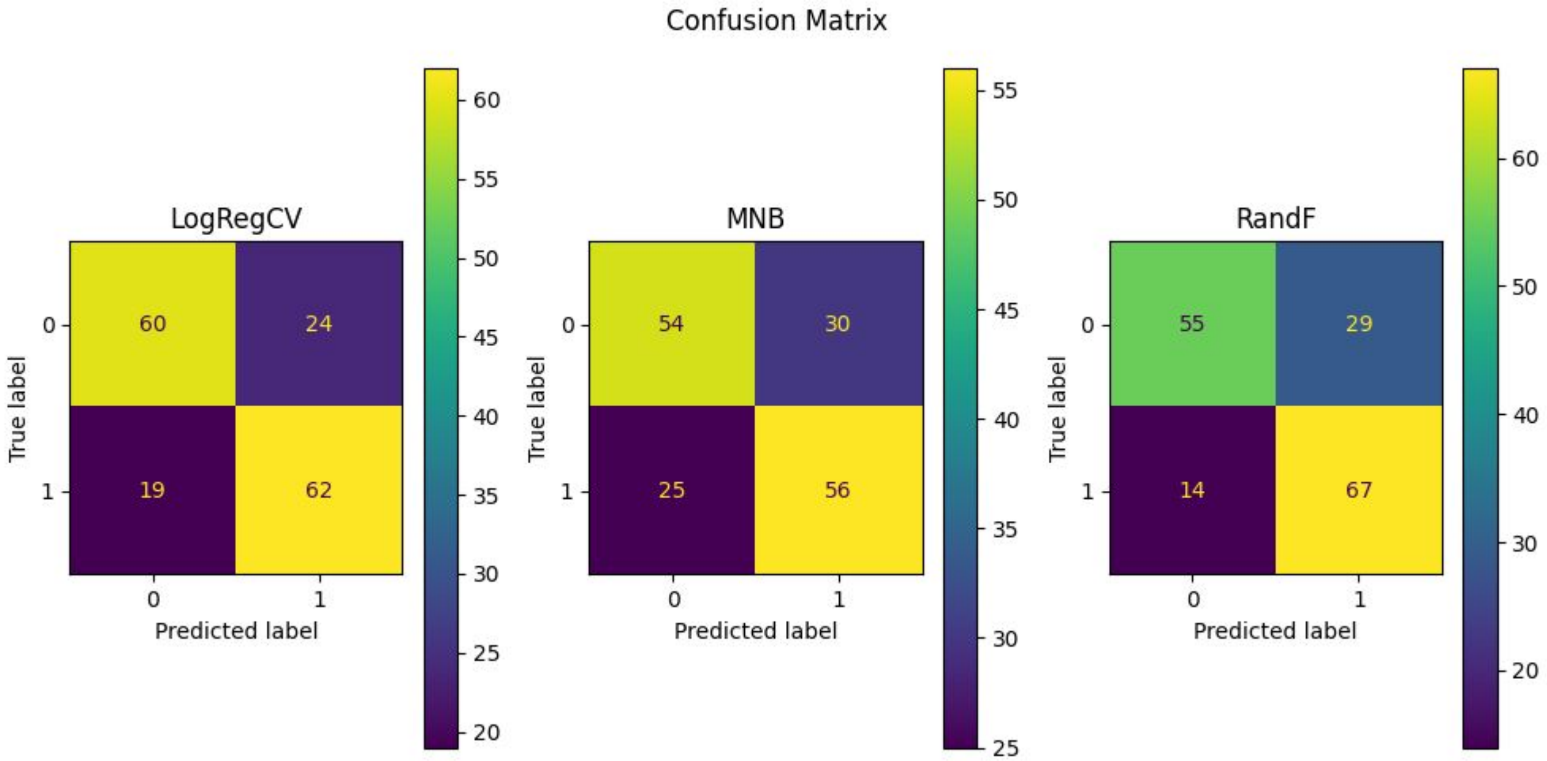
IOD_mini_project2: Classification

- Classification model comparison



IOD_mini_project2: Classification - Wrong way to sample

- Classification model comparison



IOD_mini_project2

Findings and enhancements:

1. Regression
   a. RidgeCV and LassoCV have higher evaluation period because of CV to find optimal alpha
   b. RidgeCV in particular has the highest evaluation time, because of innate LOO-CV, which is computationally expensive
   c. Linear models that were used have very similar predictions - possible improvements would be to use other types of regressor (non-linear)
2. Classification
   a. Q: Apply sampling strategy on full dataset or just the training dataset - if latter, it doesn’t make much sense, comparing the results to no sampling strategy. If former, why do we “alter” the testing data as well?
   b. A: Sampling should be done on the training data to solve the imbalance issue and not the testing data. In application, the test data should be treated as “unexpected”.
3. Overall
   a. Slides should be less technical and more story oriented, targeted at certain audiences (e.g. House buyers for house price prediction, doctors/stroke specialist working to identify high/low risk patients)
   b. Missing EDA insights (e.g. Houses that are expensive are generally closer to the ocean, high risk stroke patients are generally older etc)