ML Assignment II – Report - Hackathon

IMT2019085, IMT2019081

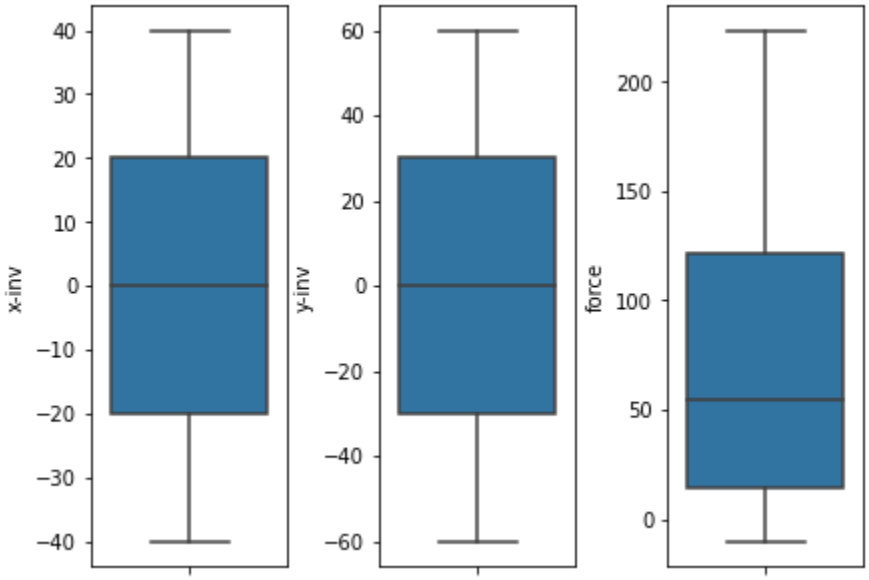
SUKHAMJOT SINGH, SHIVANSH SETHI

Team Name – NHK?

Regression Task – Dataset – Coordinates of location. We needed to predict the thrust force.

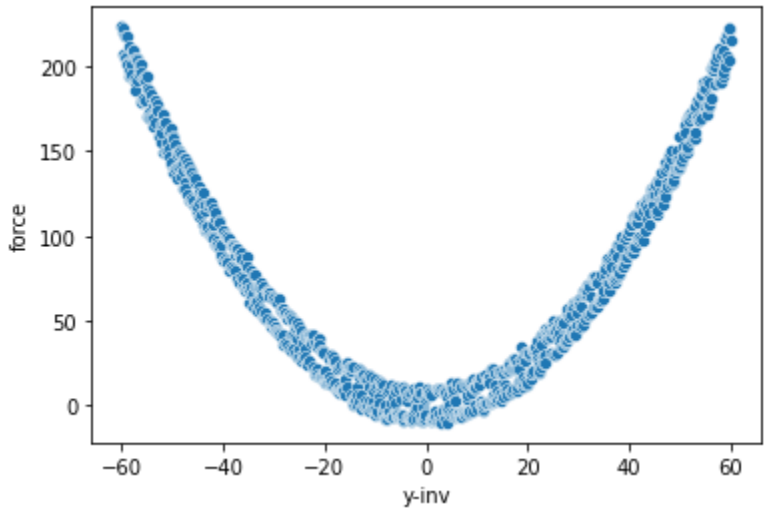
Data Preprocessing:

- There we no null values in the dataset.
- No duplicate rows were found.
- Final shape of the training dataset after preprocessing was (1000,27).
- There we no outliers observed.

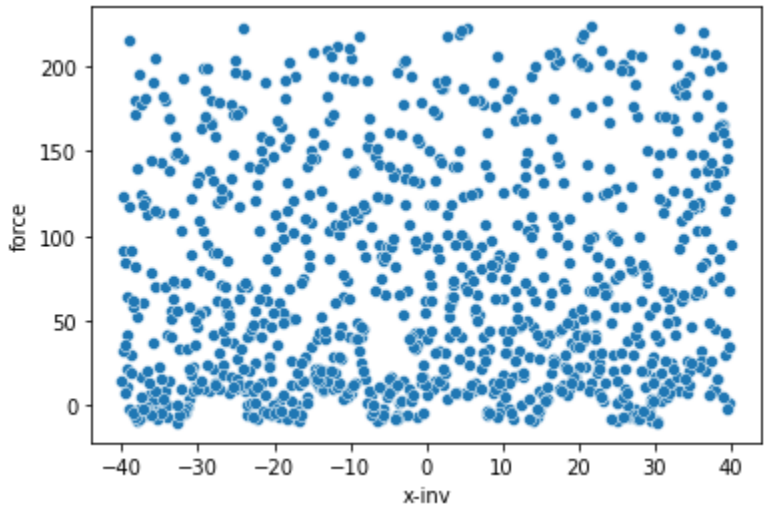


- Splitting the data into train and test datasets – Train size = 0.8.
- Polynomial features were added.

‘Y_inv’ = Degree 2.



‘X-inv’ = Degree 23



Top 3 approaches: -

1. Linear Regression model without regularization was used to predict the thrust force (with polynomial features).

   [MSE = approx. 400]

2. Linear Regression model with L1 regularization (Lasso) with polynomial features.

   [MSE = approx. 38]

3. Linear Regression model with L2 regularization (Ridge) with polynomial features.

   [MSE = approx. 8] [Minimum Error observed.]

Conclusions – Regression

- Mean Squared Error was used to evaluate the methods – Closed Form, Gradient Descent, Newton's Method.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- Parameters like Polynomial degrees and hyperparameters of regularization models were tuned by trial and error and observations to attain close to minimum error.
- We observe that the Ridge Regularization for Linear regression is the most successful in the given dataset.

Classification Task

We had to classify the data into two classes ‘poor’ and ‘rich’. Dataset- Census dataset.

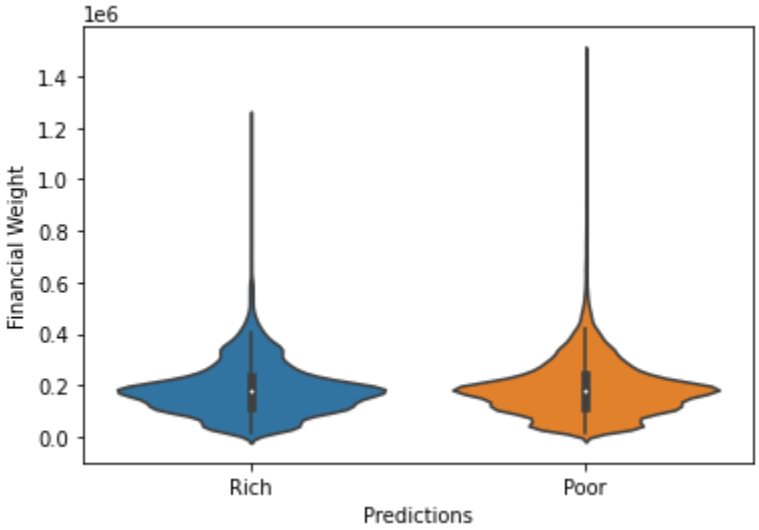
Data preprocessing -

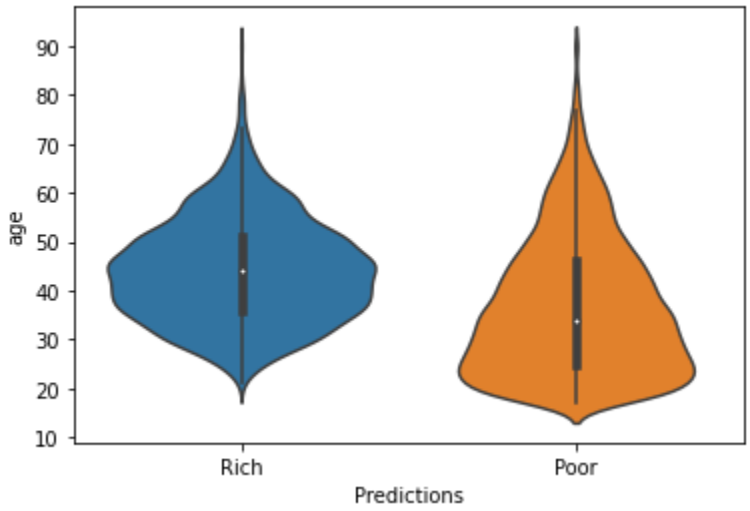
- There were ‘?’ in place of null values in some features which were replaced by nan values. Nan values for Categorical features were replaced by the most frequent category among respective categories.
- There were many categorical columns –

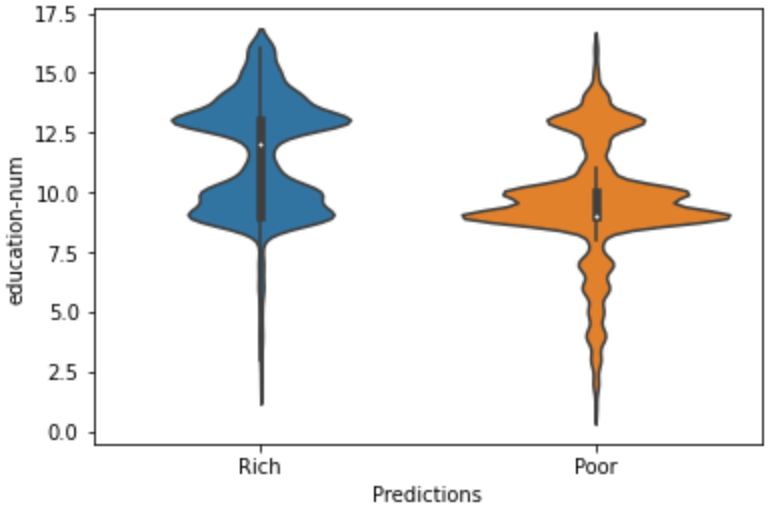
```
['Working Section', 'education', 'Marriage Status', 'occupation',
'Relationship Status', 'Skin Color', 'Gender', 'Country']
```

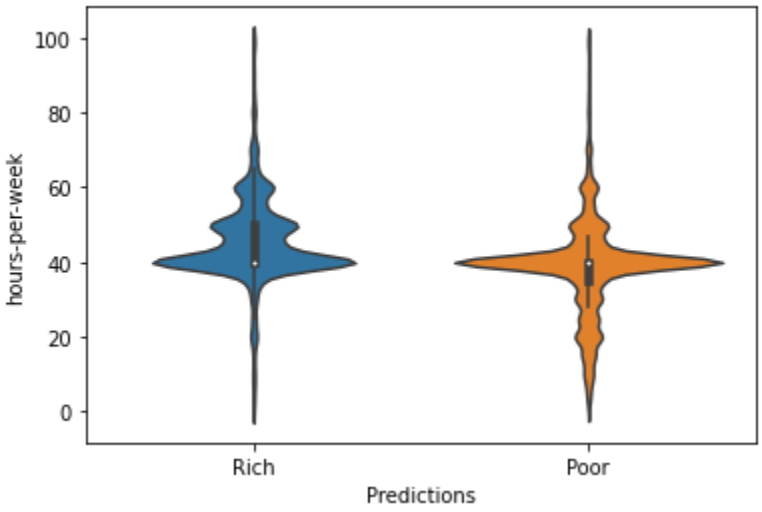
- One Hot Encoding was done on those features.
- Prediction column for rich and poor was label encoded.
- As the category United States dominated the 'Country' feature, we categorized all other countries to 'other'.

- Exploratory Data Analysis.
- Violin plots between Prediction and numerical features.

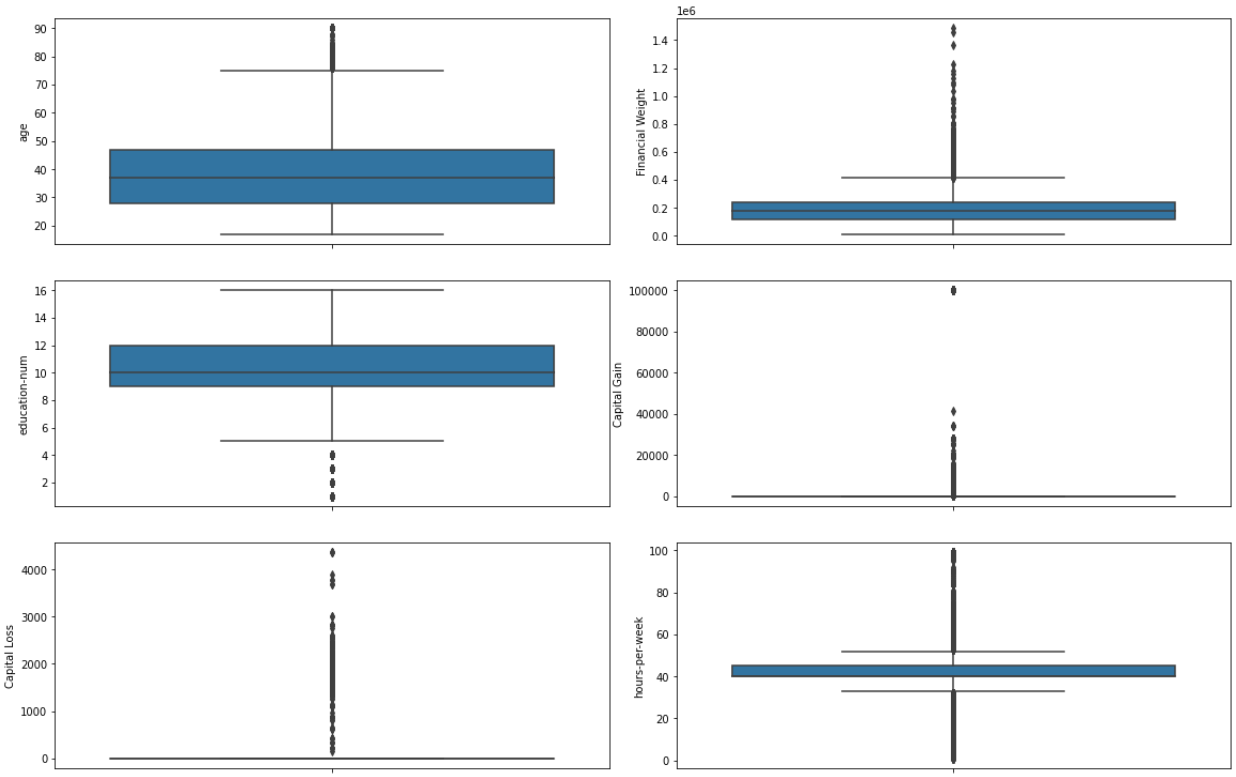




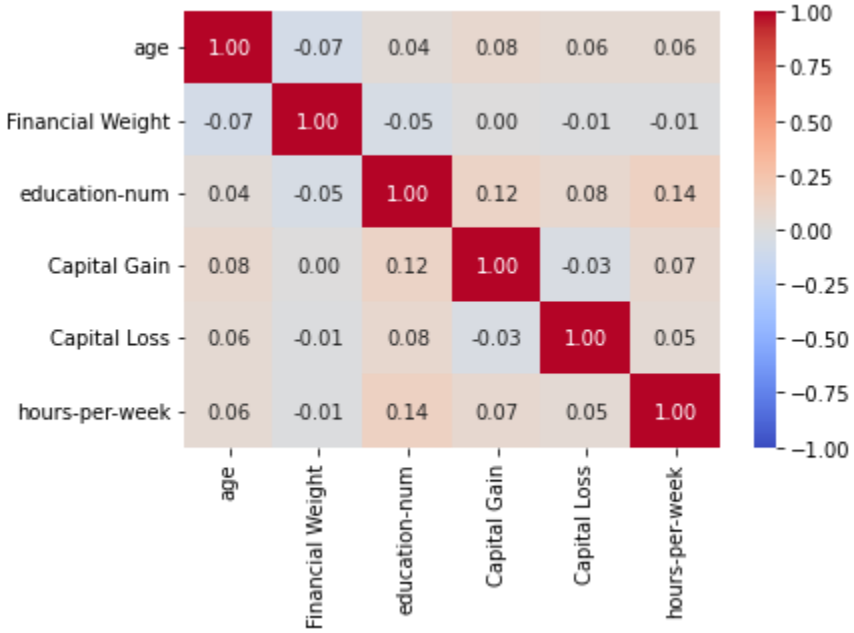

17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
education-num
Rich
Poor
Predictions


100
80
60
40
20
0
hours-per-week
Rich
Poor
Predictions

- Checking for outliers using boxplots



- Checking for correlation among numerical features



- Interactions were added between the features increasing the dimensionality of the dataset.
- Principal Component Analysis was used to decrease the dimensionality of the dataset.
- Standard Scaler was used to standardize the data.
- Final shape of data set after separating the feature to predict = (22792, 100)
- Splitting the data into train and test datasets – Train size = 0.8.

Top 3 approaches: -

1. Naïve Bayes Classifier was used to predict.
Auc roc score = [approx. 82]

2. Logistic Regression was used to predict.
Auc roc score = [approx. 88]

3. Stacking classifier [Logistic Regression, Gaussian Naïve Bayes , SVM, Bernoulli Naïve Bayes] was used to predict.
Auc roc score = [approx. 91]

Conclusions – Regression Task

- Area under the ROC curve was used to evaluate the methods.
- Parameters like dimensionality and hyperparameters of models were tuned by trial and error and observations to attain close to maximum score for the methods.
- We observe that the Stacking classifier performed best among all other methods.