Machine Learning Project Documentation

Linear Regression Model

a. General Information about the dataset:

- Our dataset called “Insurance.csv” contained insurance details for various patients. The fields were as follows: age, sex, bmi, number of children, smoker or not, region, charges.

- The dataset contained 1338 rows. The train/test split was 80/20 respectively.

b. Implementation details:

Feature extraction: There were total of 7 features extracted, 4 numerical features and 3 non numerical. The numerical fields were: age, bmi, no. of children, charges. The non numerical fields were: sex, smoker, region. We also applied data augmentation to create a new feature: **smoker_bmi**. This is because charges more likely to be high for someone who has a high bmi and is a smoker.

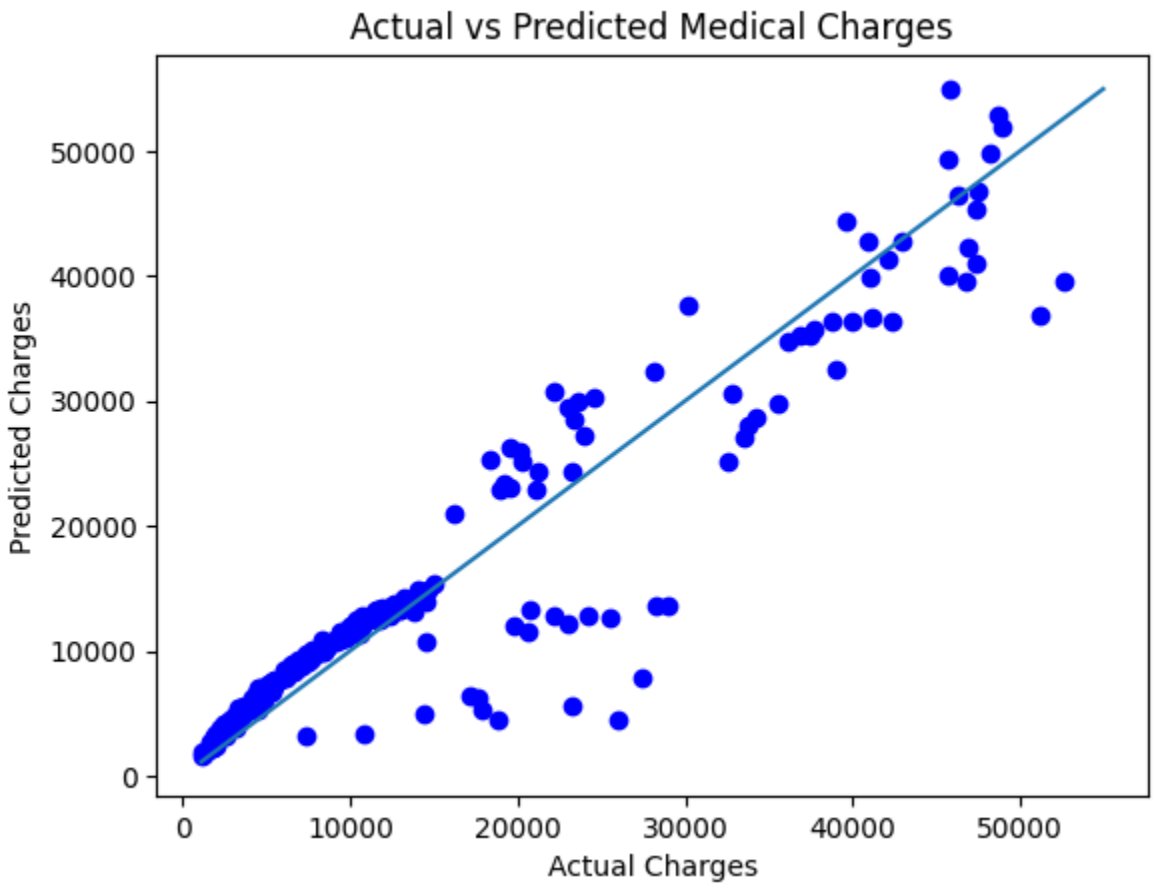
Dimensions of resulted features: We used label encoder for the sex, smoker fields and one hot encoder for the region field. Therefore the dimensions are: 6 numerical fields(including sex and

smoker) + 4 categorical fields(for region; because it has 4 unique values) = 10 features.

Cross Validation: Cross validation was not used in training the model, only a train test split.

Hyperparameters: no hyperparameters were tuned.

c. Results



Mean Squared Error: 0.133630063949074 53

R^2 Score: 0.8796076642739292

KNN Model

d. General Information about the dataset:

- Our dataset called “Insurance.csv” contained insurance details for various patients. The fields were as follows: age, sex, bmi, number of children, smoker or not, region, charges.

- The dataset contained 1338 rows. The train/test split was 80/20 respectively.

e. Implementation details:

Feature extraction: There were total of 7 features extracted, 4 numerical features and 3 non numerical. The numerical fields were: age, bmi, no. of children, charges. The non numerical fields were: sex, smoker, region. We also applied data augmentation to create a new feature: **smoker_bmi**. This is because charges more likely to be high for someone who has a high bmi and is a smoker.

Dimensions of resulted features: We used label encoder for the sex, smoker fields and one hot encoder for the region field.

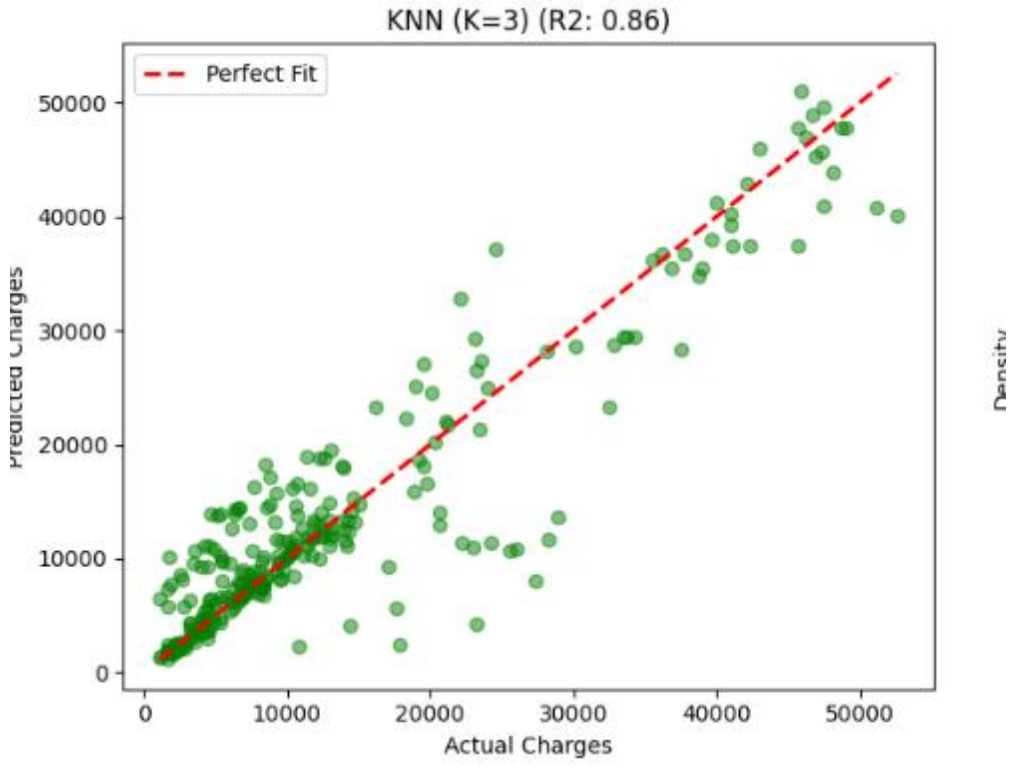
Therefore the dimensions are: 6 numerical fields(including sex and smoker) + 4 categorical fields(for region; because it has 4 unique values) = 10 features.

Cross Validation: Cross validation was not used in training the model, only a train test split.

Hyperparameters: the only hyperparameter tuned was the value of K which is the number of neighbors used in the model. It was calculated dynamically by a for loop. For k=1 to k=20, we calculated the R^2 score for each k, then picked the k corresponding to the best score.

f. Results

Graph:



KNN Mean Squared Error: 0.160708406793599 16

KNN R^2 Score: 0.8552117697700844

Logistic Regression Model

g. General Information about the dataset:

- Our dataset called “Cell Images” contained 2 classes of cell images: Malaria Parasitized cells and Uninfected cells.

- The dataset contained 13,780 Uninfected cell images and 13,780 Parasitized cell images to make a total of 27,560 images.

h. Implementation details:

Feature extraction:

- The mean and variation of Hue, Saturation, Value (HSV) color of each image
- Edge density of each image
- Blue, green, red (BGR) color ratio of each image
- Histogram representing intensity distribution across each image

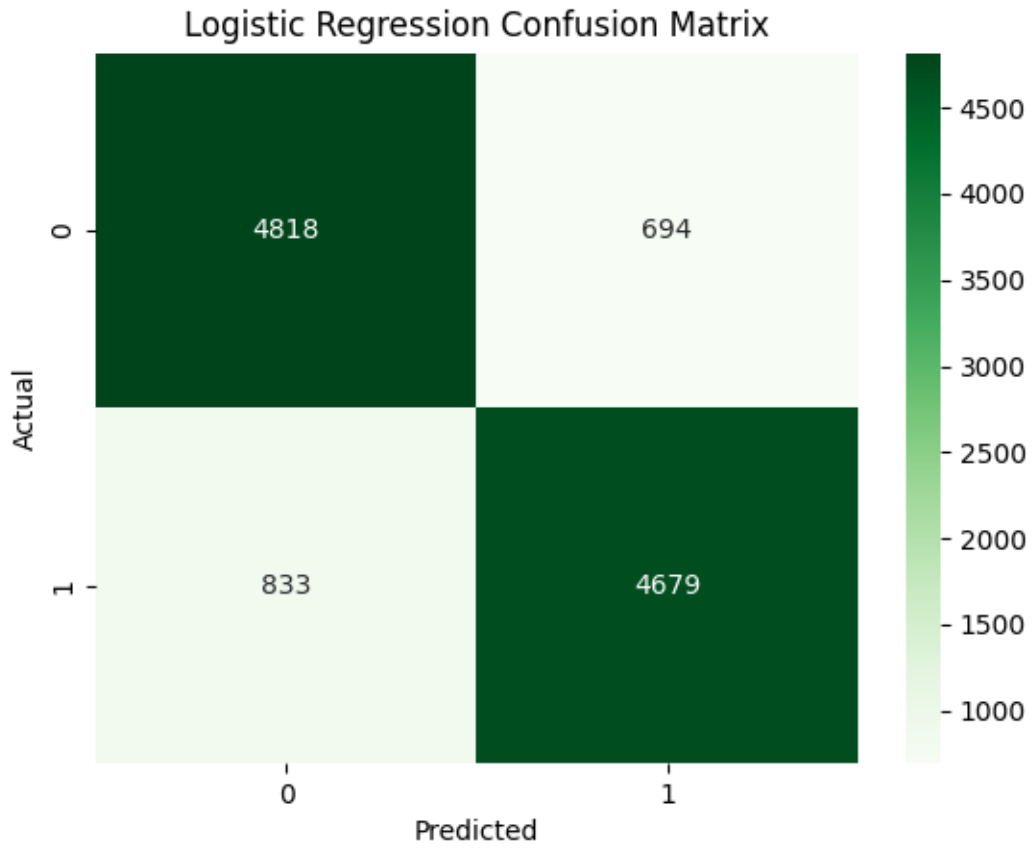
Dimensions of resulted features: Each feature vector has about 13 features total.

Cross Validation: Cross validation was not used.

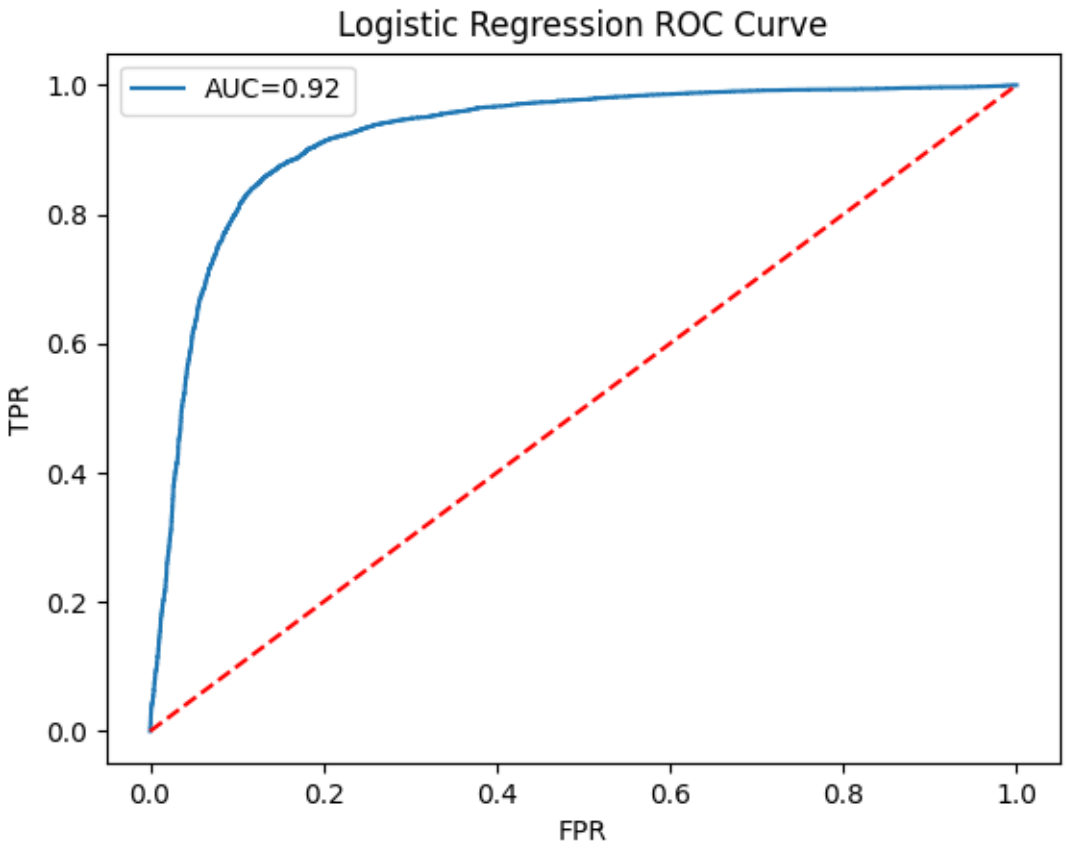
Hyperparameters: PCA n_components: set to 400.
Logistic regression solver=”lbfgs”, max_iter=1000

i. Results

Confusion Matrix



ROC Curve



Logistic Regression Accuracy: 0.87

K-Means Regression Model

j. General Information about the dataset:

- Our dataset called “Cell Images” contained 2 classes of cell images: Malaria Parasitized cells and Uninfected cells.

- The dataset contained 13,780 Uninfected cell images and 13,780 Parasitized cell images to make a total of 27,560 images.

k. Implementation details:

Feature extraction:

- The mean and variation of Hue, Saturation, Value (HSV) color of each image
- Edge density of each image
- Blue, green, red (BGR) color ratio of each image
- Histogram representing intensity distribution across each image

Dimensions of resulted features: Each feature vector has about 13 features total.

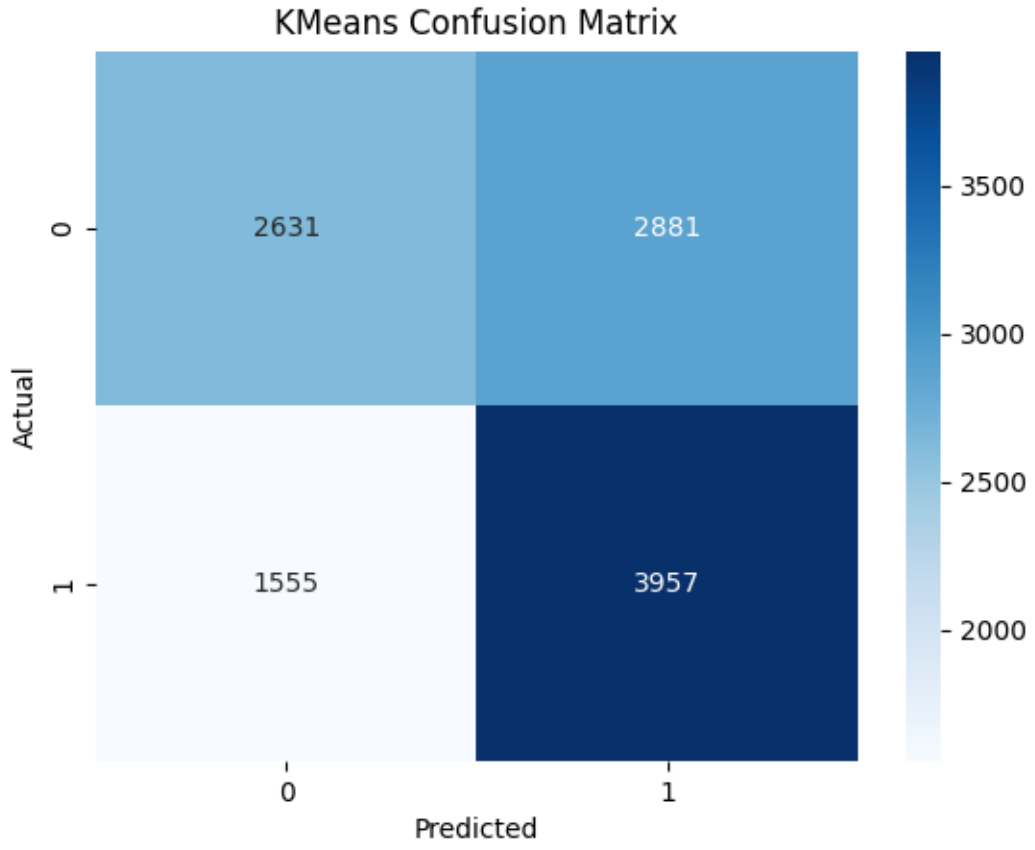
Cross Validation: Cross validation was not used.

Hyperparameters: n_clusters was set to 2.

n_init was set to 20.

l. Results

Confusion Matrix



K-Means Accuracy: 0.61