Predict cancer mortality rates for US counties.

Case Study 1

By: Anand mohan
18BCS6218

CONTENTS

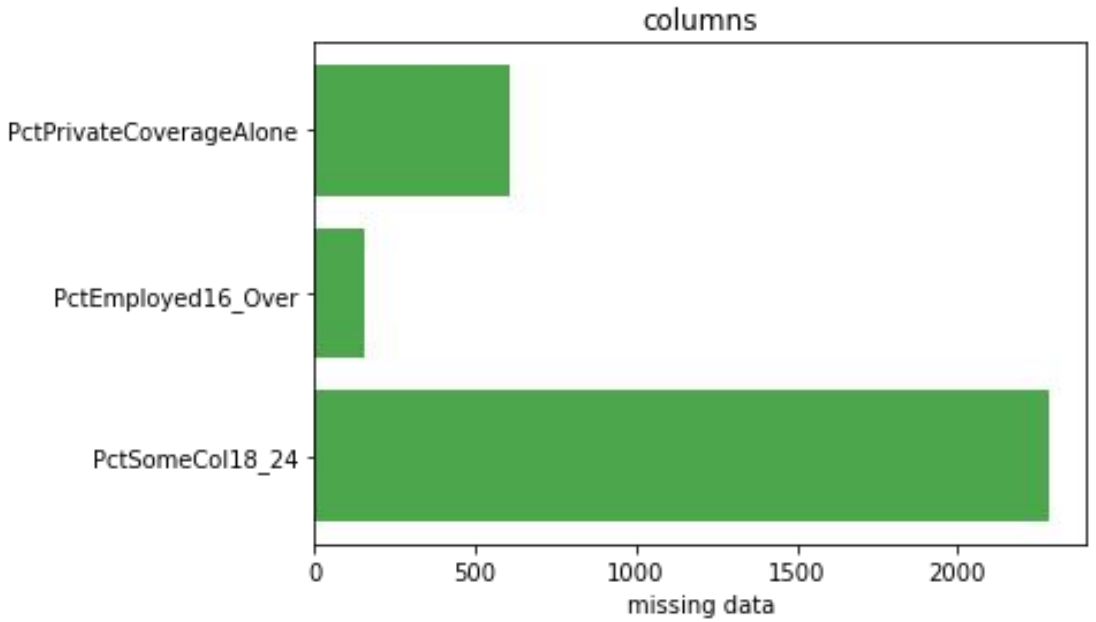
Part 01

Data Preparation

Information about Dataset

Missing Values

Out of 34 columns 3 columns have missing values



Part 02

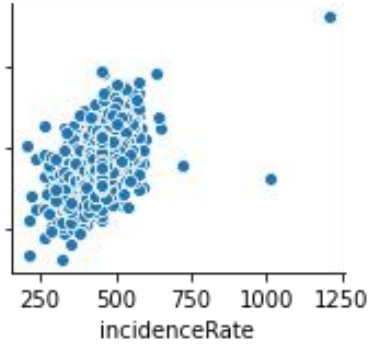
Data Exploration

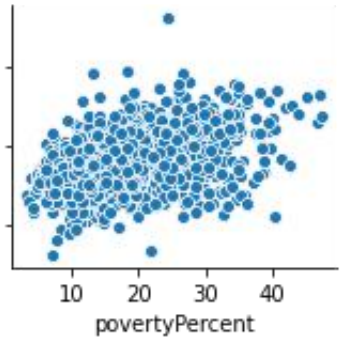
Relationship between Data

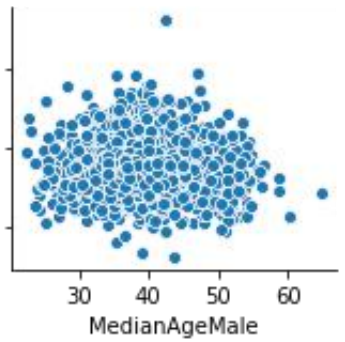
Most columns dont have a linear relationship with the target columns but some have

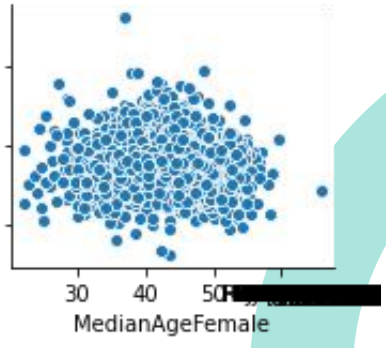
Categorical Data after some processing showed relatioship with the target column

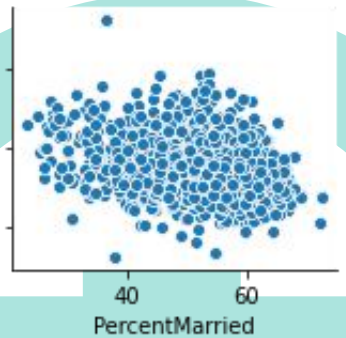
Numeric Graphs

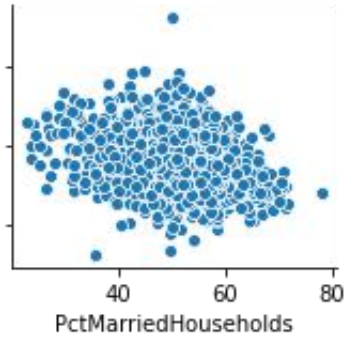


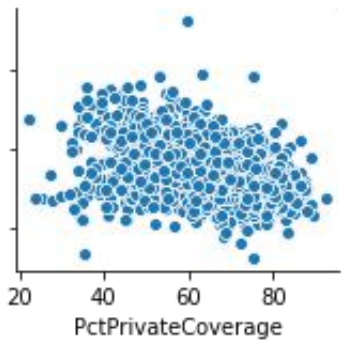


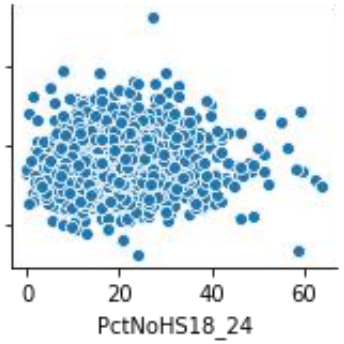


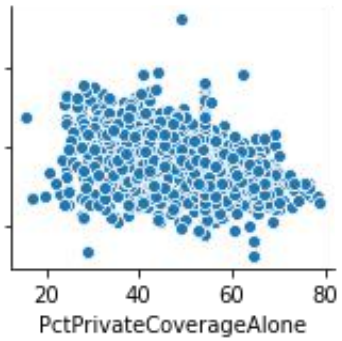


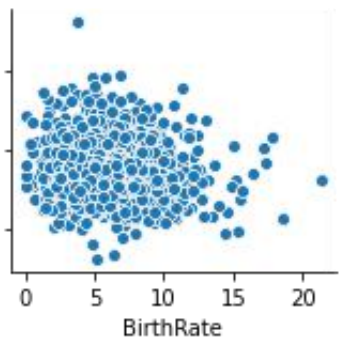


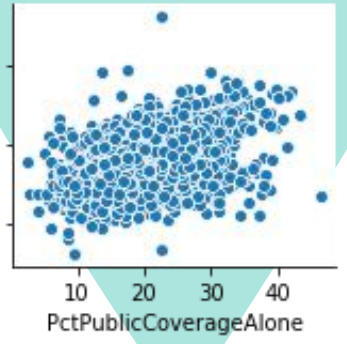






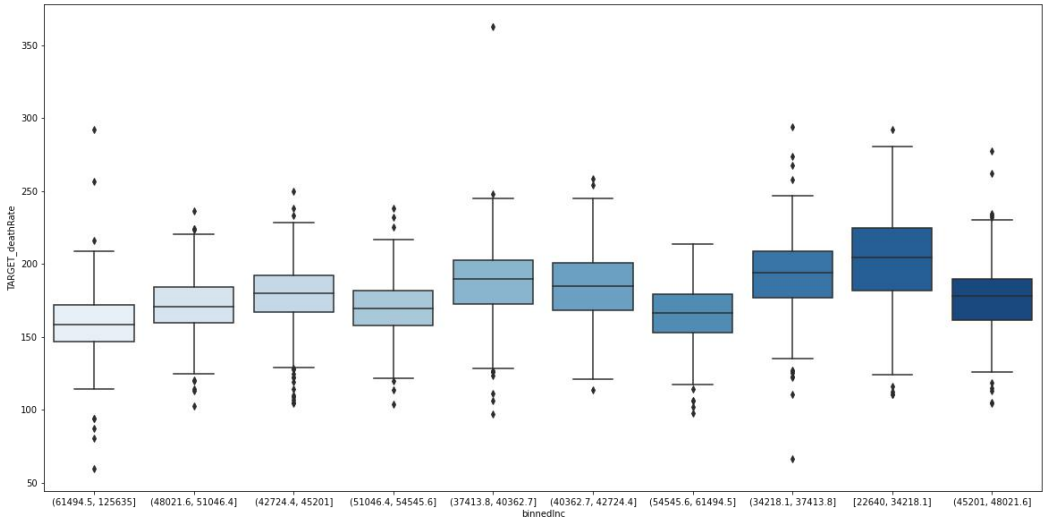






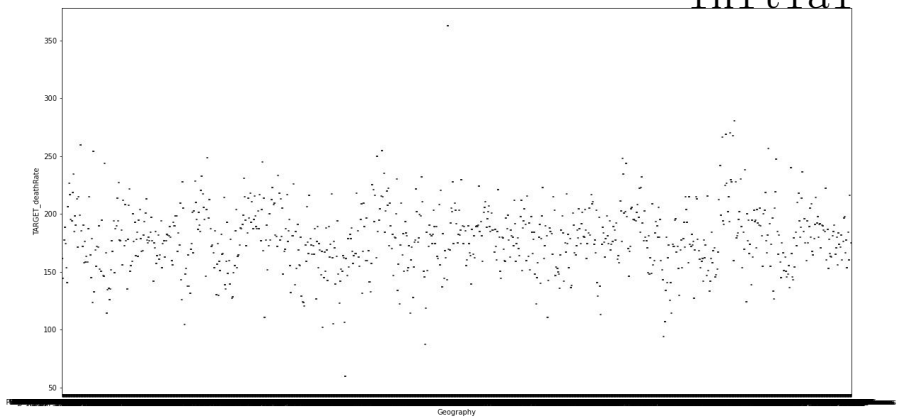
Categorical Relationship

Median income per capita binned by decile vs Target Deathrate

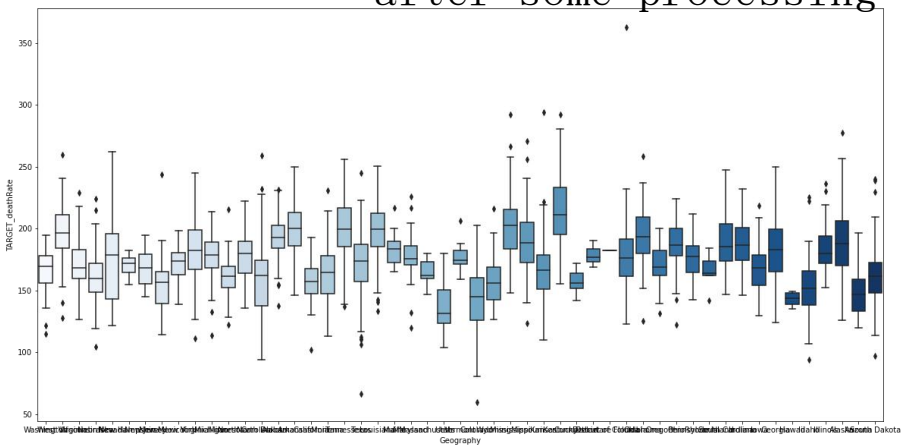


Geography vs Target Deathrate

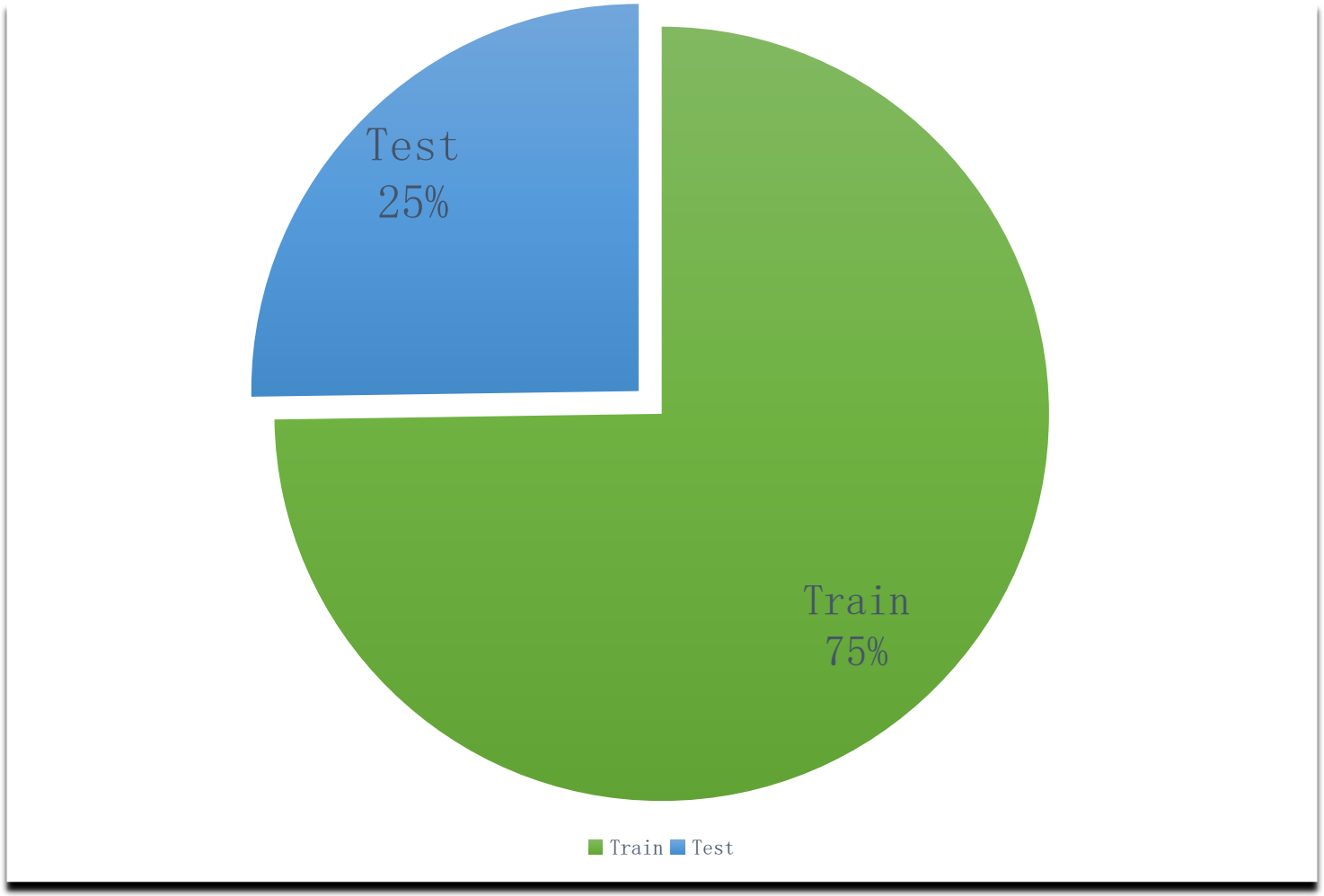
initial



after some processing



Split

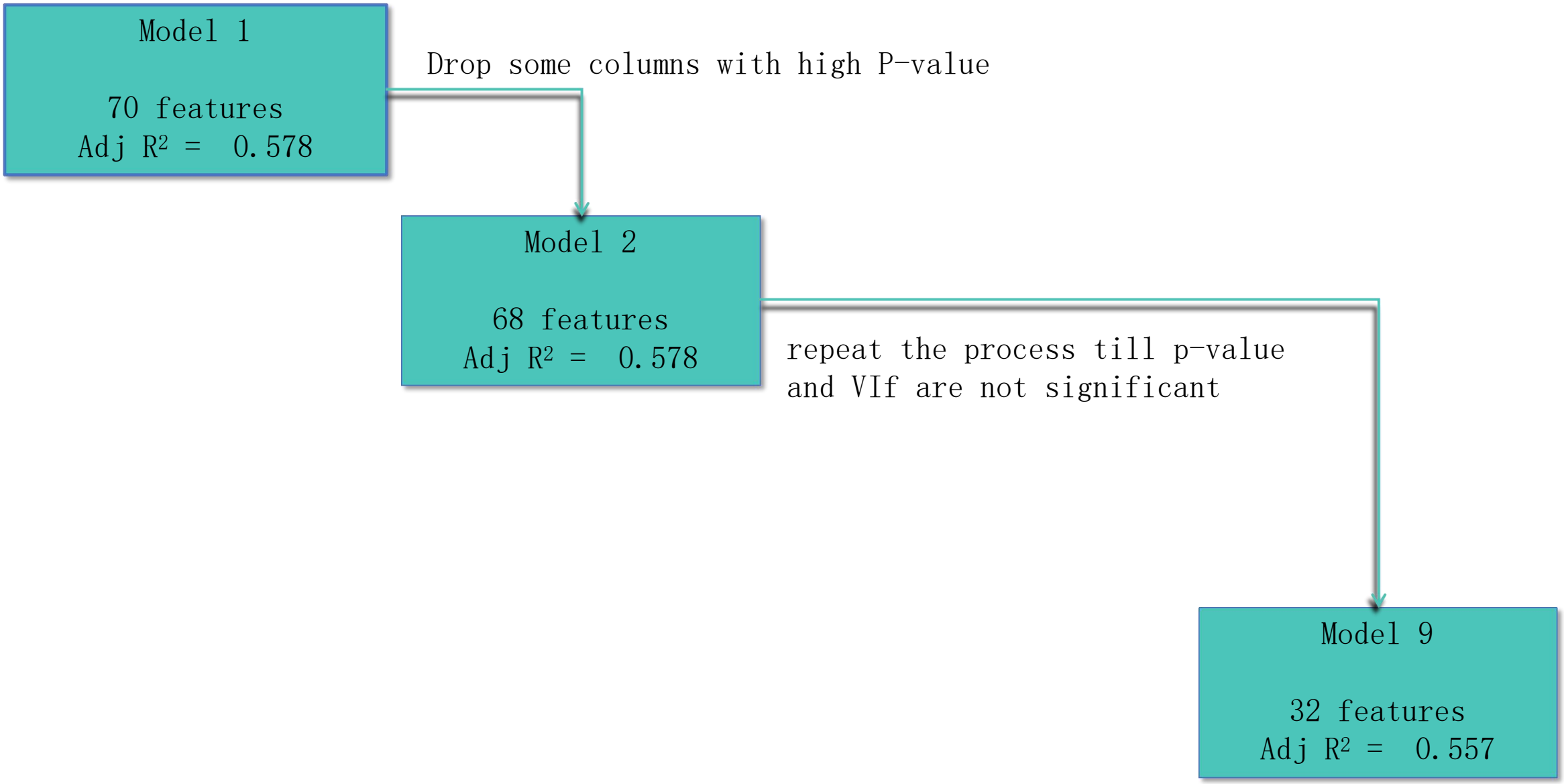


Data Distribution

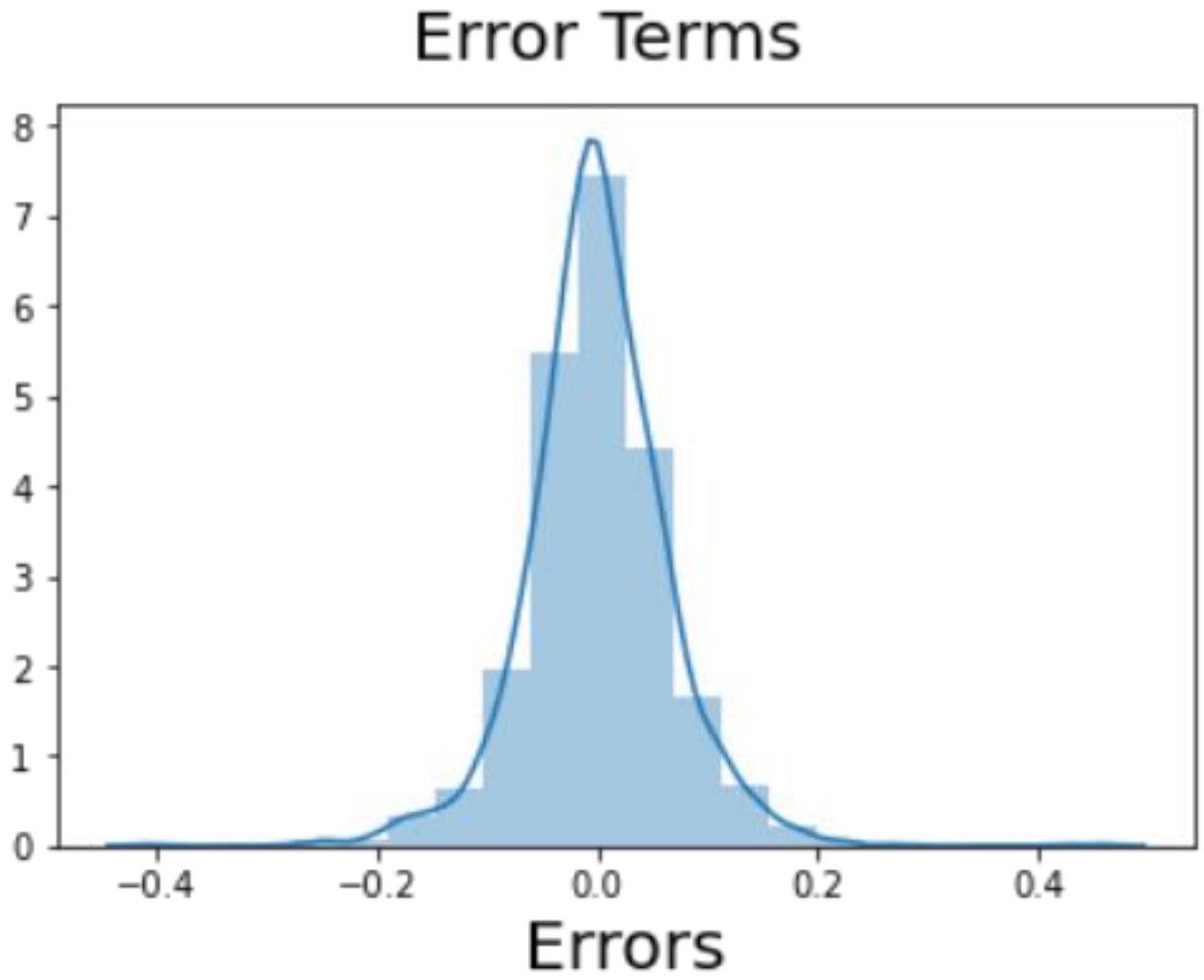
Part 03

Training

Model Build

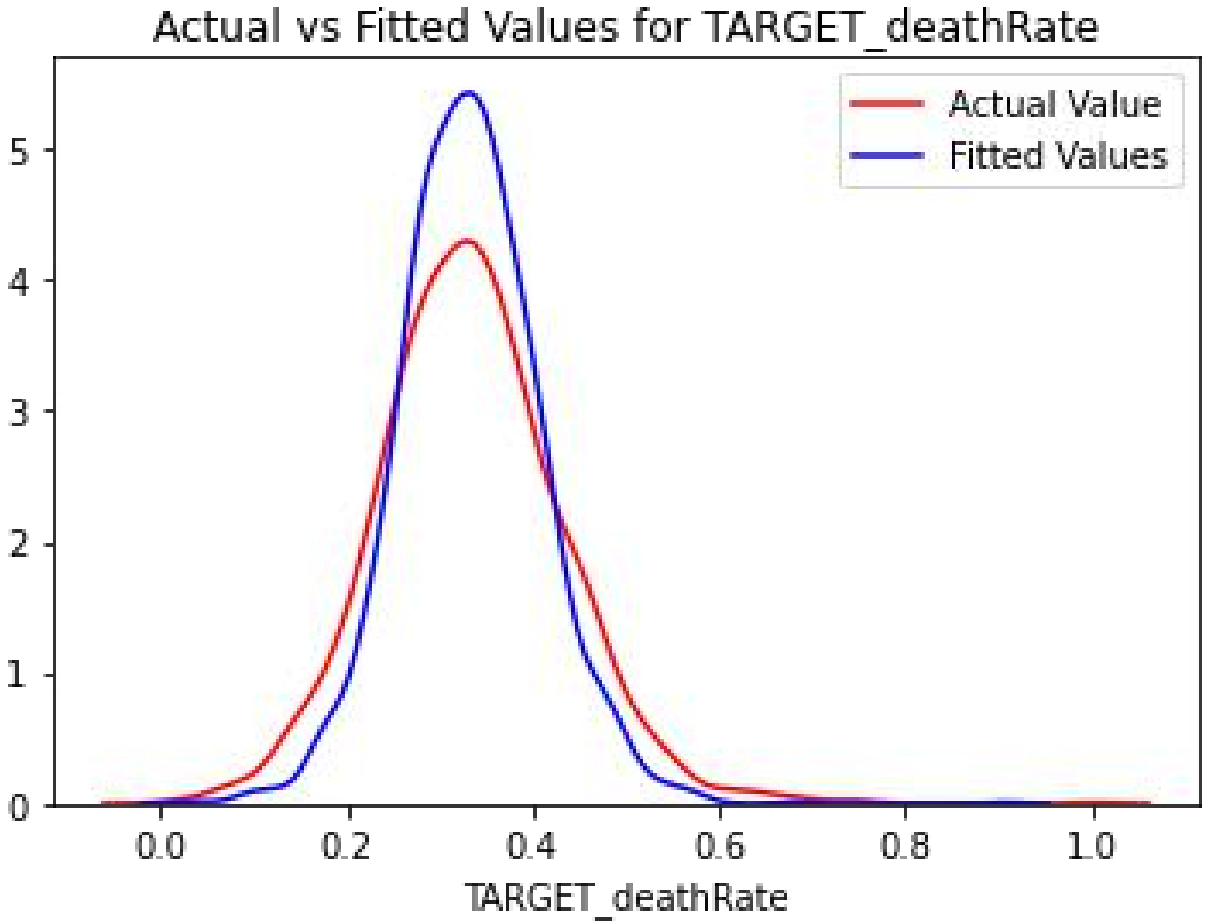

Model 1
70 features
Adj R^2 = 0.578
Drop some columns with high P-value
Model 2
68 features
Adj R^2 = 0.578
repeat the process till p-value
and VIf are not significant
Model 9
32 features
Adj R^2 = 0.557

Residual Analysis



The Error terms are normally distributed

Residual Analysis

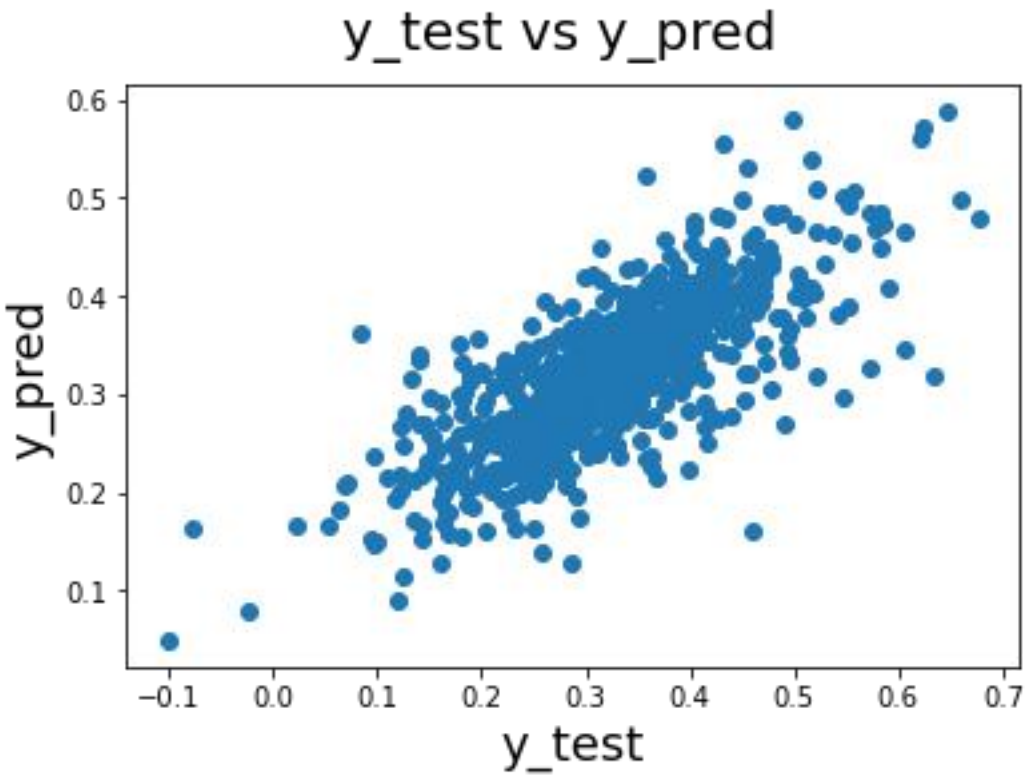


We can see that the fitted values are reasonably close to the actual values, since the two distributions overlap a bit. However, there is definitely some room for improvement.

Part 04

Testing

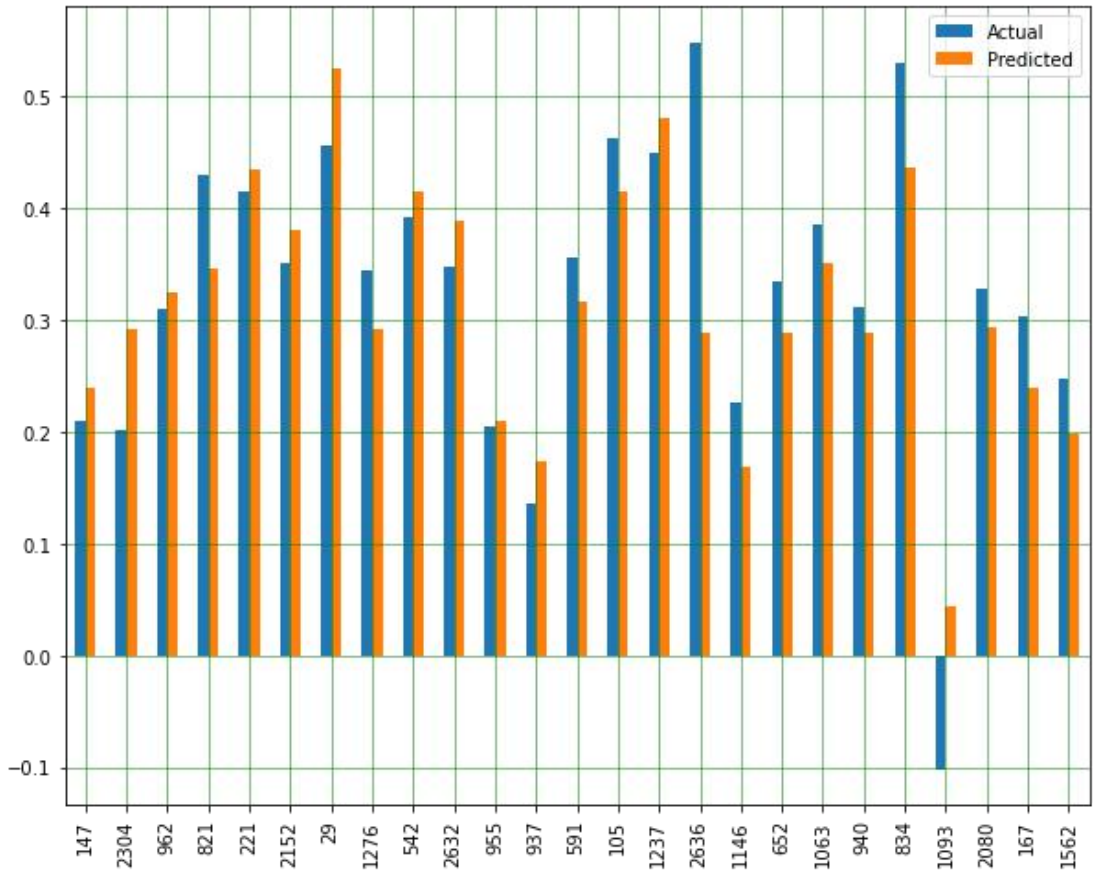
Model Evaluation

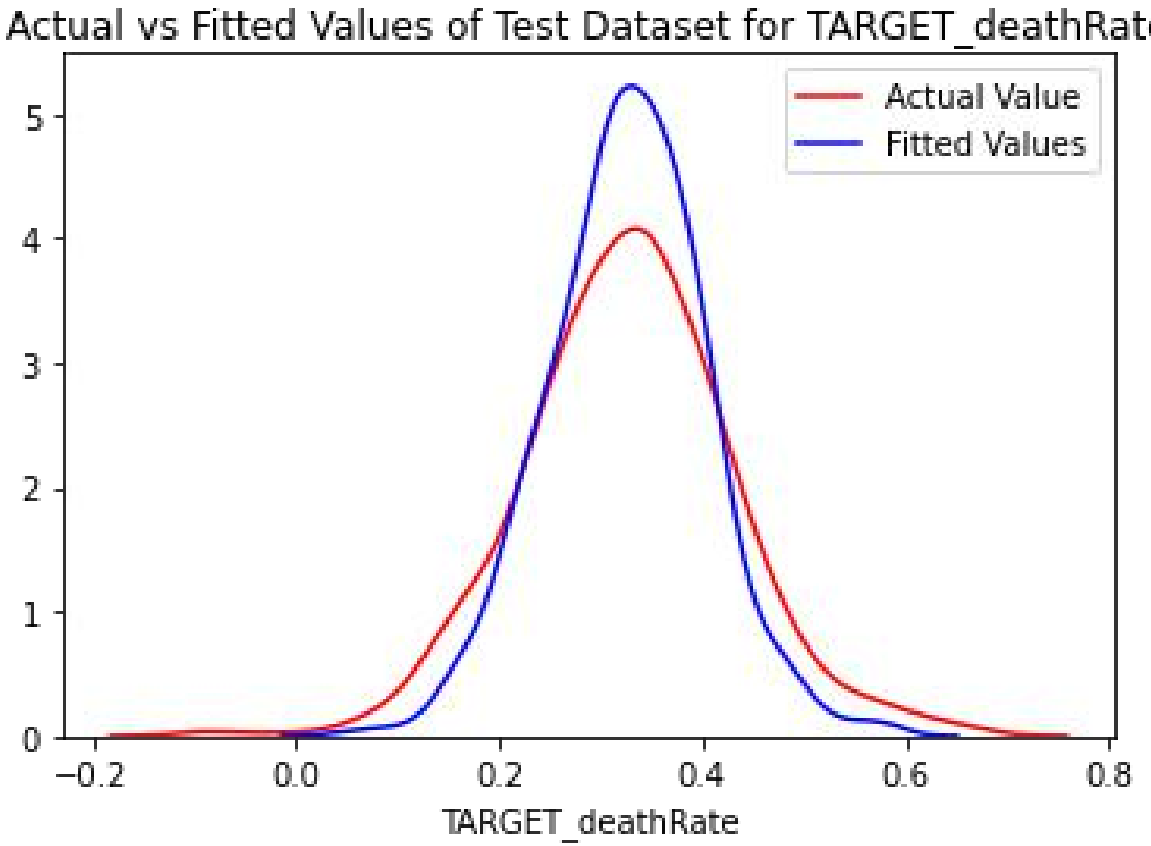


Target = incidenceRate*0.65+MedianAgeFemale*-0.06+PctHS18_24* 0.05+PctHS25_Over* 0.04+PctBachDeg25_Over*-0.16+PctUnemployed16_Over*0.08+PctPublicCoverageAlone*0.06+PctOtherRace*-0.07 +PctMarriedHouseholds*-0.07+BirthRate *-0.06+avg Income bwn[22640, 34218.1]*0.02+Alaska*0.07+Arizona* -0.05+Arkansas *0.04+California *-0.04+Colorado* -0.05+Georgia* -0.02+Hawaii* -0.07+Idaho*-0.04+Iowa * -0.02+Kentucky *0.03+Montana* -0.04+New Mexico*-0.05+New York* -0.05+North Carolina* -0.02+Oklahoma*0.05+Oregon*-0.02+Pennsylvania*-0.03+Tennessee* 0.02+Utah*-0.09+Virginia*0.02

Model Evaluation

Mean Absolute Error: 0.050336134511656 7
Mean Squared Error: 0.0047229149770521995
Root Mean Squared Error: 0.0687234674405 4901





Part 05

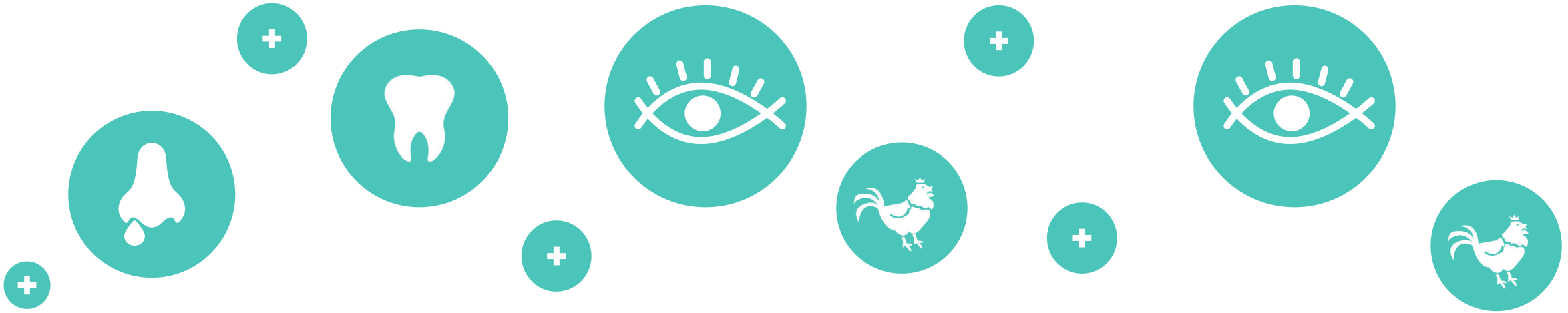
Conclusion

Conclude

Overall we have a decent model, but we also acknowledge that we could do better.

We have a couple of options:

Add new features (avgAnnCount/avgDeathsPerYear/PctWhite etc.)
Build a non-linear model
Remove or Transform Outliers



Thank You

All the Documents are uploaded on GitHub: github.com/tenoob/Machine-Learning/tree/master/sem5/CaseStudy1