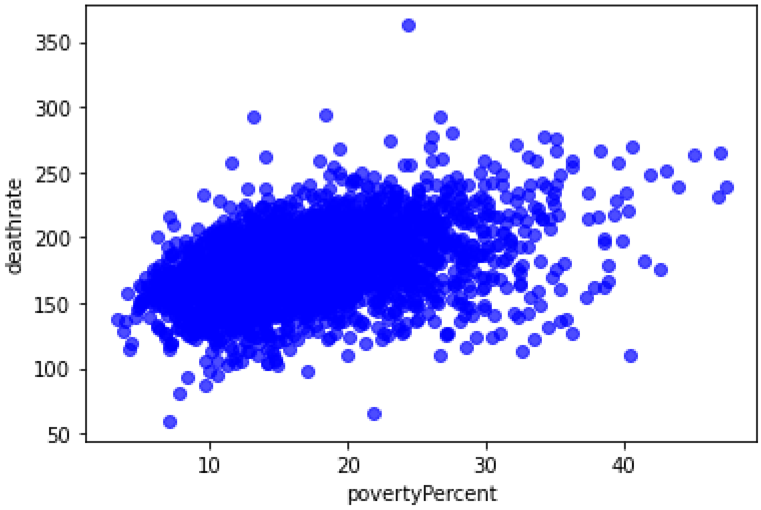
CSIS 2390
Term Project Report
By
Prabjot Singh Wilkhu
300357845

Introduction and Discovery

It has been observed that cancer is one of the leading factors of the death of a human. But in recent times the health care system has taken a great leap in terms of curing cancer. Apart from the improvement in the medical technology and health care systems there are other various factors that lead to mortality due to cancer. Some of these factors are geological, financial, age, government schemes, education level of different age groups of a particular city etc.

These factors primarily determine whether a person with cancer has access to ways to get treatment to cure cancer. For example the below scatter plot depicts that the mortality due to cancer is more in cities with higher poverty rate which implies that people with poor financial condition has less access to proper health care facilities.



So, such factors should be taken into consideration because just the advancements in the technology to cure cancer is of no use if common citizens of the country have no access to them.

Dataset

The dataset has the collection of values consisting of features of the demographics of various cities of the United States of America. The purpose of the dataset is to provide attributes of the demographics like mean age, mean household income, insurance policies of an individual, level of education etc and the mortality rate due to cancer. This dataset helps to determine the correlation between the mortality rate and the other features associated with the demographics of a particular region. The dataset has been taken from the “data.world” website.

Entire dataset is numeric except the Geography and binnedInc column. Geography column contains the city and the state for the particular row. Those columns are not of much use hence, they were dropped. A few column names were renamed for better understandability.

The mean poverty percentage of a city is 16.87 and the standard is 6.40 while the minimum is 3.2. Mean percentage of paper covered under public insurance is 36.25, the minimum is 11.2 and the standard is 7.84.

Three feature selection methods were put in use before passing the dataset. These feature selection methods were Correlation based selection, Variance Threshold Selection, SelectKBest based selection. The Correlation based selection was performed by manually selecting features with high correlation with the deathrate from the correlation matrix and the dataset was named “data_corr”. The Variance Threshold feature selection was done by taking the threshold of 0.18 and the dataset was named “data_vt”. The SelectKBest based feature selection was performed by taking the score function as f_regression and k as 2. The dataset made out of it was named “data_selkBest”.

After this the dataset was split into train and test with 75 percent training data and 25 percent testing data. Apart from this feature scaling like Robust was also done and Polynomial feature transformation of degree two is also done.

The model was made to run through the combination of regressors, feature scaling and feature transformation in order to find the best predictive model out of all.

Model Planning

According to nature of the dataset the regression models will suit best in order to predict the target i.e. “deathrate”. The regression models used were as follows :

- LinearRegression
- DecisionTreeRegressor
- GradientBoostingRegressor
- RandomForestRegressor
- AdaBoostRegressor
- SGDRegressor

As the dataset had the feature which were in the correlation with the target i.e. “deathrate” the regression models could effectively be put in use to predict the target.

Instead of using a single regression model a number of different regression models were used to get deep insights of how regression models work in a particular dataset.

Dataset after feature selection methods will be passed through the scalar and then be passed through the pipeline implementation.

Model Implementation

Firstly seven lists were made to keep record of a score of a particular model with different feature transformation and feature scaling methods. These lists were made to store the regression model, r2 to store r^2 score, rmse to keep root mean square error record, fsel_list to keep the track of which feature selection method is being used, ftrans_list to keep record of feature transformation method and fscal_list to keep record of feature scaling method.

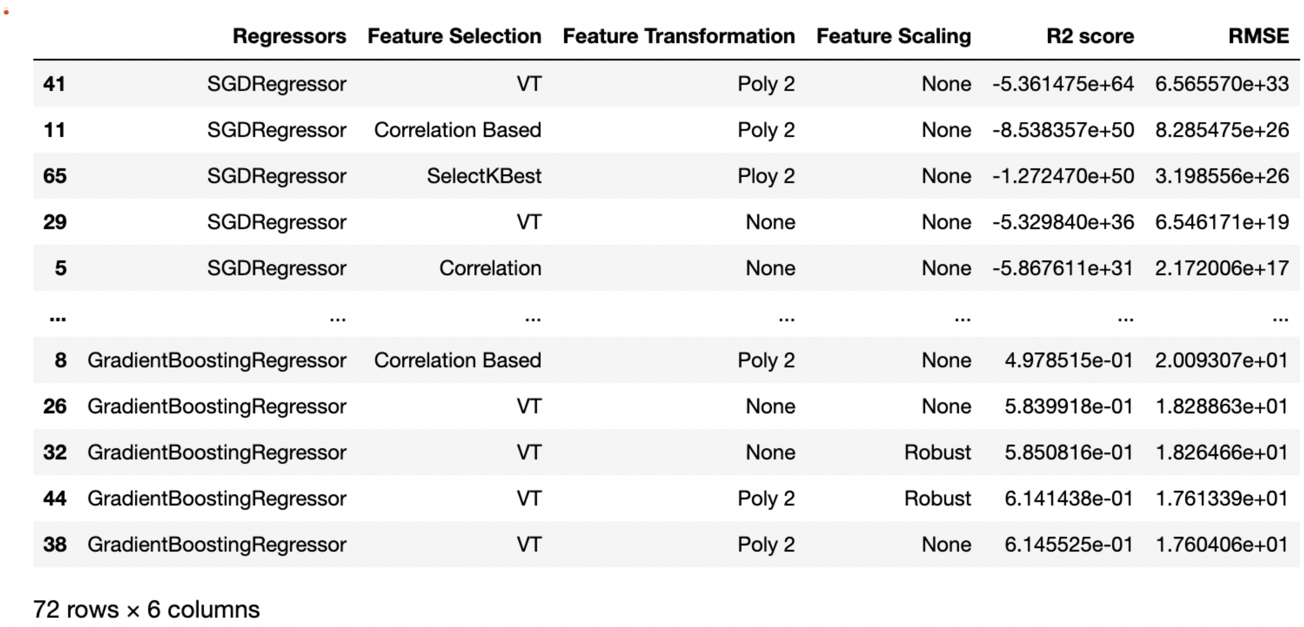
Pipeline was used to run the models through the dataset. Using a single pipeline an entire list of models ran efficiently. Apart from the regressors, scalers were also passed through the pipeline. Polynomial feature transformation was also done to test the dataset.

A for loop iterating through the models list was used and the regression models were passed using the pipeline within the loop. By using this all the planned regression models were able to process the data and give the predictions.

Earlier in model planning feature scaling was decided to be done separately before passing it through the regression pipeline but while implementation a modification was made by scaling the dataset through the pipeline itself in order to improve the efficiency.

Results Interpretation and Implications

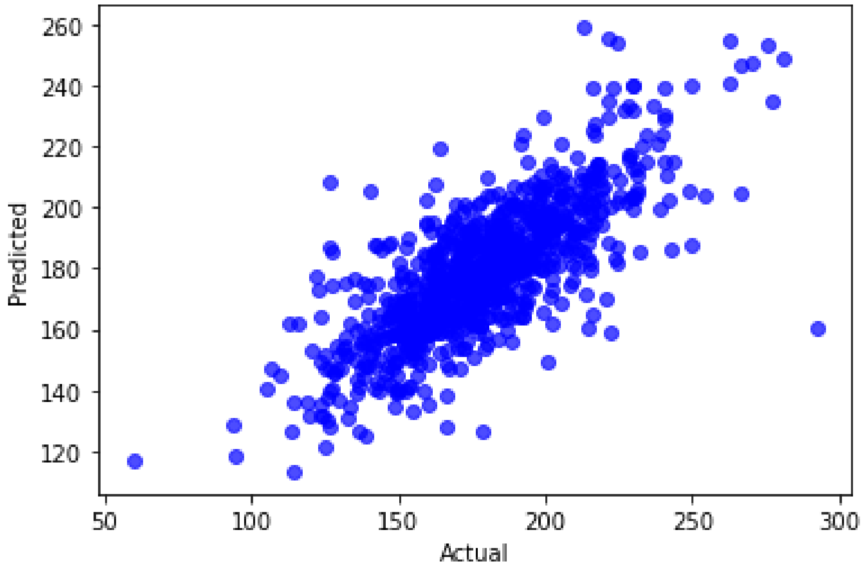
After passing the dataset with different combinations of regression models along with different selection and feature scaling and feature transformation. The result came out to be.



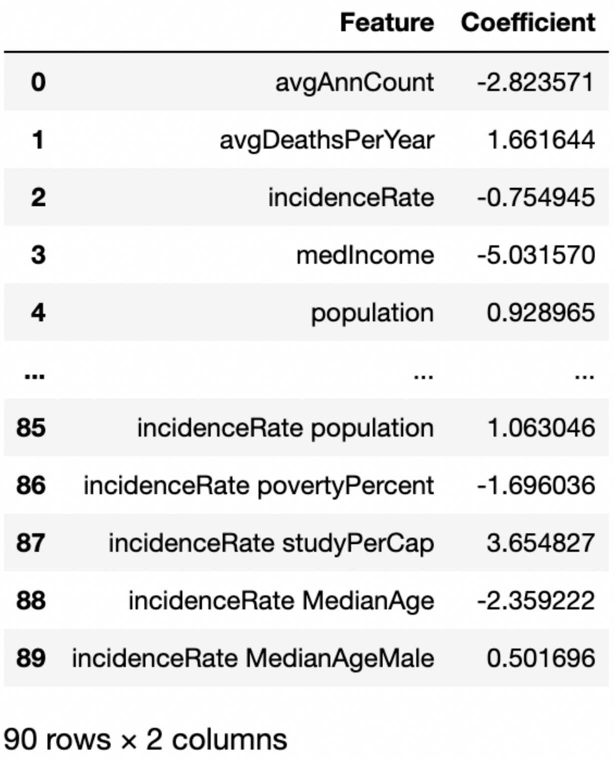
	Regressors	Feature Selection	Feature Transformation	Feature Scaling	R2 score	RMSE
41	SGDRegressor	VT	Poly 2	None	-5.361475e+64	6.565570e+33
11	SGDRegressor	Correlation Based	Poly 2	None	-8.538357e+50	8.285475e+26
65	SGDRegressor	SelectKBest	Ploy 2	None	-1.272470e+50	3.198556e+26
29	SGDRegressor	VT	None	None	-5.329840e+36	6.546171e+19
5	SGDRegressor	Correlation	None	None	-5.867611e+31	2.172006e+17
...	...	...	...	...	...	...
8	GradientBoostingRegressor	Correlation Based	Poly 2	None	4.978515e-01	2.009307e+01
26	GradientBoostingRegressor	VT	None	None	5.839918e-01	1.828863e+01
32	GradientBoostingRegressor	VT	None	Robust	5.850816e-01	1.826466e+01
44	GradientBoostingRegressor	VT	Poly 2	Robust	6.141438e-01	1.761339e+01
38	GradientBoostingRegressor	VT	Poly 2	None	6.145525e-01	1.760406e+01

72 rows × 6 columns

It has been sorted in ascending order to get the maximum R^2 score. From this table it can be observed that the best Regression model is GradientBoostingRegressor with Variance Threshold Feature Selection and Polynomial with degree two feature transformation.



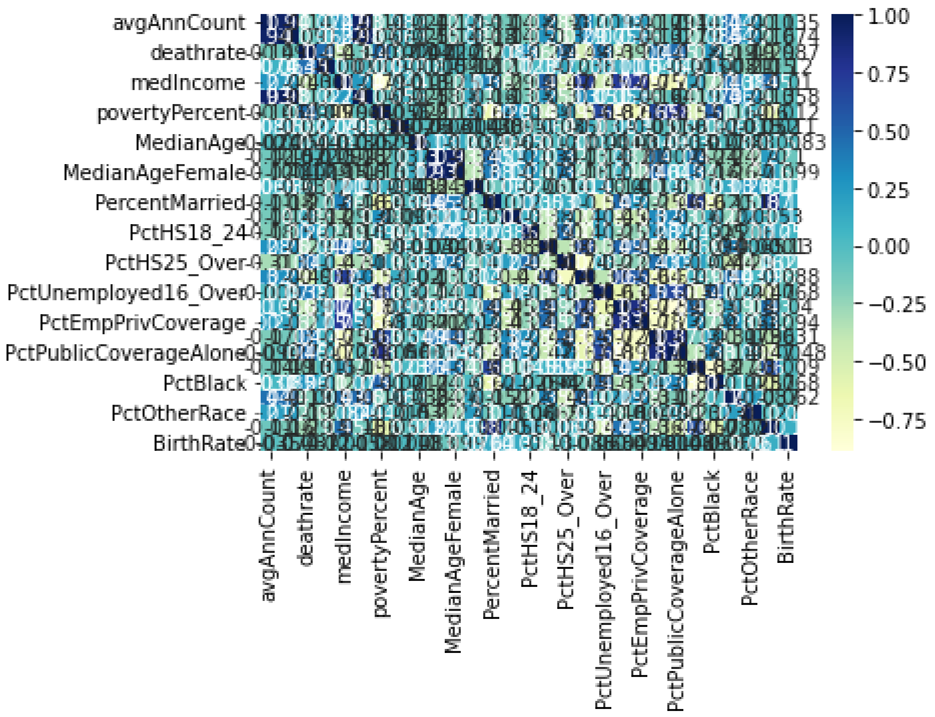
The scatter plot of y_test v y_pred depicts the good correlation hence, the model selected works well with the dataset.



	Feature	Coefficient
0	avgAnnCount	-2.823571
1	avgDeathsPerYear	1.661644
2	incidenceRate	-0.754945
3	medIncome	-5.031570
4	population	0.928965
...	...	...
85	incidenceRate population	1.063046
86	incidenceRate povertyPercent	-1.696036
87	incidenceRate studyPerCap	3.654827
88	incidenceRate MedianAge	-2.359222
89	incidenceRate MedianAgeMale	0.501696

90 rows × 2 columns

Coefficient List



It can be analysed that factors such as level of education, type of insurance, household income, Age also contribute to the mortality due to cancer. Few features of the dataset are in high correlation indicating the importance of them to be considered.

Conclusion

So the model seconds the hypothesis that the external factors are also a matter of concern when it comes to mortality due to cancer.