CAPSTONE PROJECT

CAUSALENS

ANA BEATRIZ POTJE

APPLIED DATA SCIENCE BOOTCAMP

CAMBRIDGE SPARK

SUPERVISED BY RICARDO MONTI / KIRE KOLAROSKI

LONDON, 26TH MARCH 2020

1. Introduction

Objective

The objective of this project is to compare results between four machine learning regression models implemented in Python code, and define the one which best predicts a target variable on a given dataset. SHAP (SHapley Additive exPlanations) approach is to be used to explain the feature correlations of the best model.

The best predicting model is then to be compared with CausaLens and determine which model returns best prediction results as well as the most influential features used for the target prediction.

Scope

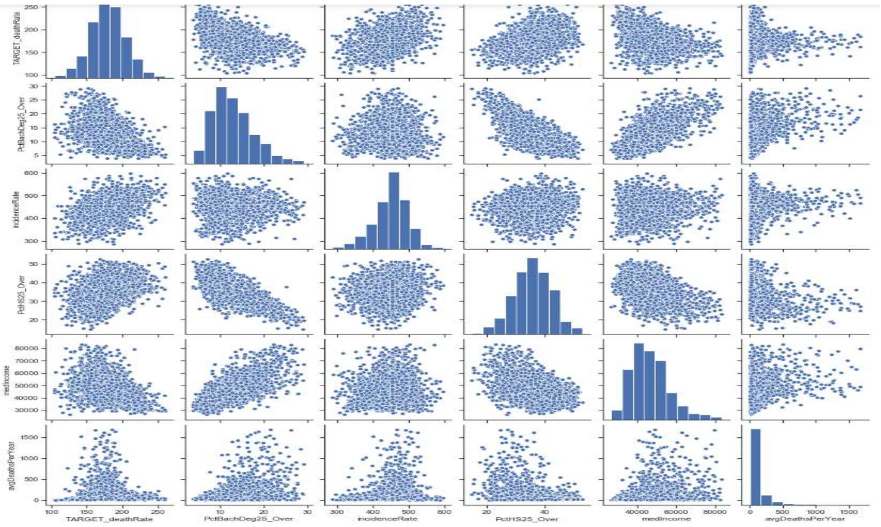
“Cancer Mortality Prediction in the US Counties” dataset was provided on the web page: https://data.world/nrippner/ols-regression-challenge . “TARGET_deathRate” – the target variable – is the mean per capita (100,000) of cancer mortalities per US county in the period between 2010 and 2016 (or 2013 Census Estimates). This dataset contains 3,047 rows and 34 features for each of the US counties for this period, which are described in the document “Capstone Project CausaLens - Ana Potje – APPENDIX”.

2. Exploratory Data Analysis

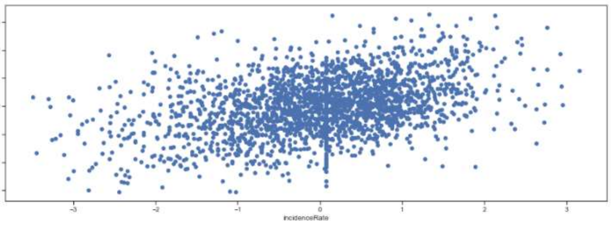
In order to deploy the best machine learning model to predict our target variable (TARGET_deathRate), the dataset must be initially prepared and unwanted issues removed. The following operations were then performed in a Jupyter Notebook (Python) environment:

- Categorical features mapped to integer values (“Geography”);
- Dropped other categorical features not required (“binnedInc”);
- Filled in null values with the value “0”;
- Removed Outliers from the dataset;
- Scaled the dataset, changing its mean to value 0 and standard deviation of 1.

A Pair Plot of the features which seems to be more related to rate of cancer mortality is below:



A Scatter Plot of an important feature (Incident Rate) against the Target DeathRate is presented below:



3. Model Description / Results

The dataset was split in the following subsets: Train (64% of the data), Validation (16% of the data) and Test (20% of the data). The metrics used to evaluate and compare the models were:

a) **L2 norm** – calculates the distance between the real and predicted values of our target variable and
b) **R-Squared or Coefficient of Determination** – used to measure how much of the variation in outcome can be explained by the variation in the independent variables.

The regression models used are below:

Regression Model	Package / Class	Parameters	L2 Norm	R-Squared
Linear Regression (baseline)	Scikit-learn / LinearRegression	default values	260.79	**0.4257**

Decision Tree Regression	Scikit-learn / DecisionTreeRegressor, GridSearchCV	criterion = mse,
		max_depth = 6,
		max_leaf_nodes = 100,
		min_samples_leaf = 20,
		min_samples_split = 3
			336.70	**0.3233**

Random Forest Regression	Scikit-learn / RandomForestRegressor**,** GridSearchCV	criterion = mse,
		max_depth = 8,
		max_leaf_nodes = 100,
		min_samples_leaf = 10,
		min_samples_leaf = 5
			260.03	**0.4660**

SVR	Scikit-learn / SVR	kernel = rbf,
		C = 0.5
			262.18	**0.4511**

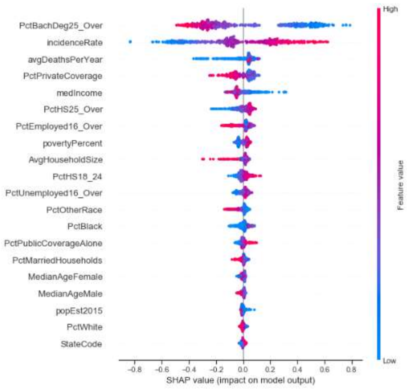
4. Best Model Analysis

The best performing model was the Random Forest Regression, with highest R-Squared value of 0.4660. Its Feature Importance table (top features) is presented below:

	Variable	Importance
1	PctBachDeg25_Over	0.32
2	incidenceRate	0.22
3	avgDeathsPerYear	0.05
4	medIncome	0.04

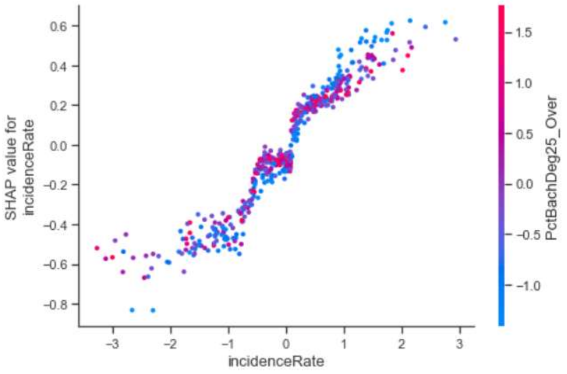
SHAP Analysis

The features that are most important for this model are represented in the plot below, sorted by the sum of SHAP value magnitudes over all samples. The SHAP values show the distribution of the impacts each feature has on the model output. E.g., high values of PctBachDeg25_Over lowers the predicted target Death Rate.



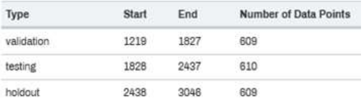
SHAP Dependency Plot

The plot below shows there is an approximately linear and positive trend between incidenceRate and the target variable, and it interacts with the feature PctBachDeg25_Over frequently.



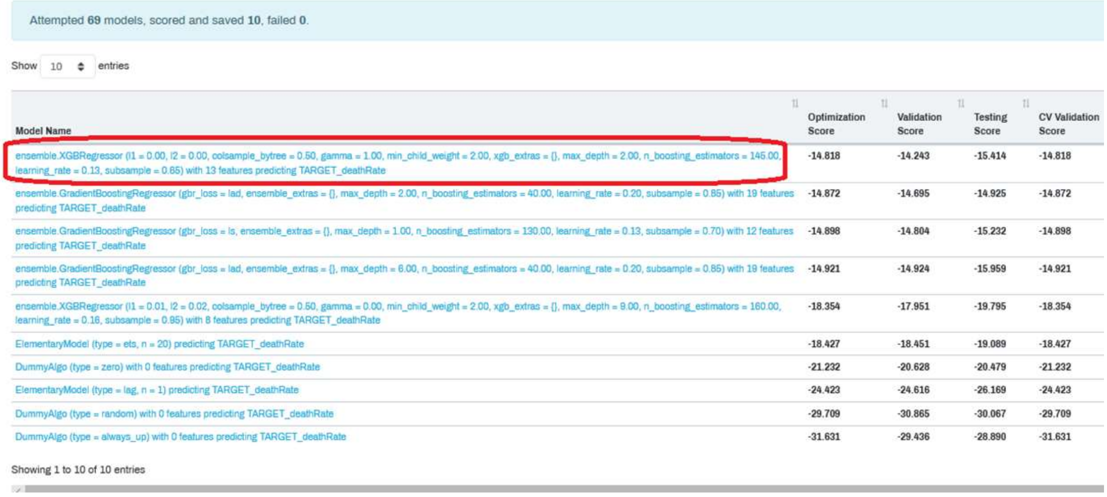
5. CausaLens Results

CausaLens platform was then configured to run on the same dataset. Parameters were set in the tool, such as target variable, maximum time/ maximum number of trials, input features, scoring metrics, etc. The data was divided in: Train (40% of data), Validation (20% of data), Test (20% of data) and Holdout (20% of data).



Type	Start	End	Number of Data Points
validation	1219	1827	609
testing	1828	2437	610
holdout	2438	3046	609

After running the tool, 69 models were generated and the top performant ones saved (table below).



Attempted **69** models, scored and saved **10**, failed **0**.

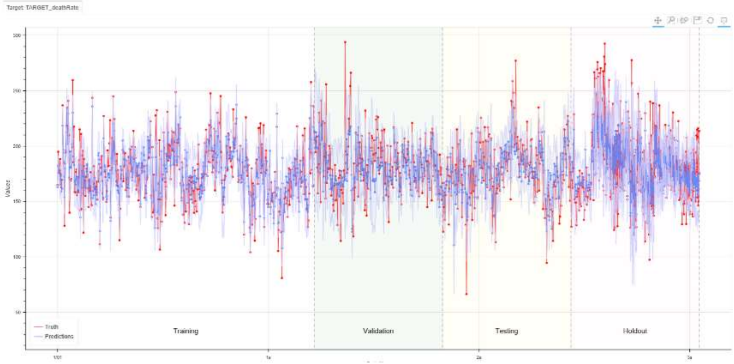
Show 10 entries

Model Name	Optimization Score	Validation Score	Testing Score	CV Validation Score
ensemble.XGBRegressor (l1 = 0.00, l2 = 0.00, colsample_bytree = 0.50, gamma = 1.00, min_child_weight = 2.00, xgb_extras = {}, max_depth = 2.00, n_boosting_estimators = 145.00, learning_rate = 0.13, subsample = 0.65) with 13 features predicting TARGET_deathRate	-14.818	-14.243	-15.414	-14.818
ensemble.GradientBoostingRegressor (gbr_loss = lad, ensemble_extras = {}, max_depth = 2.00, n_boosting_estimators = 40.00, learning_rate = 0.20, subsample = 0.85) with 19 features predicting TARGET_deathRate	-14.872	-14.695	-14.925	-14.872
ensemble.GradientBoostingRegressor (gbr_loss = ls, ensemble_extras = {}, max_depth = 1.00, n_boosting_estimators = 130.00, learning_rate = 0.13, subsample = 0.70) with 12 features predicting TARGET_deathRate	-14.898	-14.804	-15.232	-14.898
ensemble.GradientBoostingRegressor (gbr_loss = lad, ensemble_extras = {}, max_depth = 6.00, n_boosting_estimators = 40.00, learning_rate = 0.20, subsample = 0.85) with 19 features predicting TARGET_deathRate	-14.921	-14.924	-15.959	-14.921
ensemble.XGBRegressor (l1 = 0.01, l2 = 0.02, colsample_bytree = 0.50, gamma = 0.00, min_child_weight = 2.00, xgb_extras = {}, max_depth = 9.00, n_boosting_estimators = 160.00, learning_rate = 0.16, subsample = 0.95) with 8 features predicting TARGET_deathRate	-18.354	-17.951	-19.795	-18.354
ElementaryModel (type = ets, n = 20) predicting TARGET_deathRate	-18.427	-18.451	-19.089	-18.427
DummyAlgo (type = zero) with 0 features predicting TARGET_deathRate	-21.232	-20.628	-20.479	-21.232
ElementaryModel (type = lag, n = 1) predicting TARGET_deathRate	-24.423	-24.616	-26.169	-24.423
DummyAlgo (type = random) with 0 features predicting TARGET_deathRate	-29.709	-30.865	-30.067	-29.709
DummyAlgo (type = always_up) with 0 features predicting TARGET_deathRate	-31.631	-29.436	-28.890	-31.631

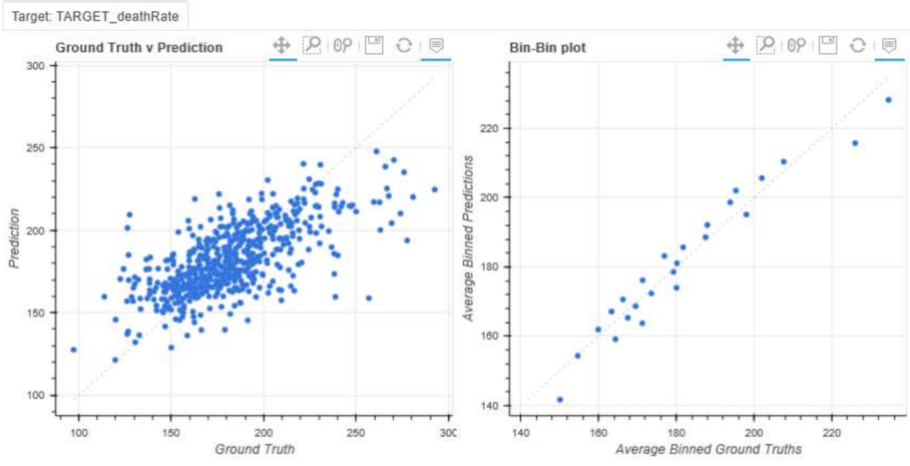
Showing 1 to 10 of 10 entries

The best score was achieved with model ensemble.XGBRegressor, using 13 features and the parameters: l1 = 0.00, l2 = 0.00, colsample_bytree = 0.50, gamma = 1.00, min_child_weight = 2.00, xgb_extras = {}, max_depth = 2.00, n_boosting_estimators = 145.00, learning_rate = 0.13 and subsample = 0.65.

Holdout Full Performance

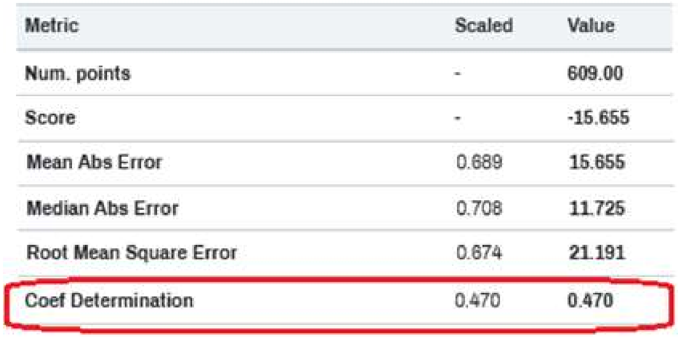


Holdout Prediction X Ground Truth Performance



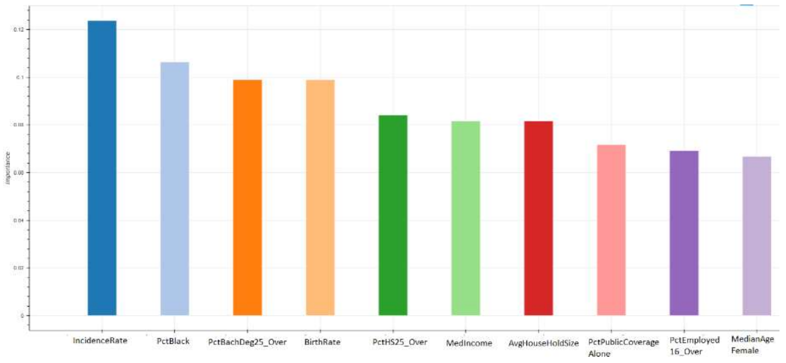
Data sorted by prediction values and averaged into 25 bins.

Performance Metrics

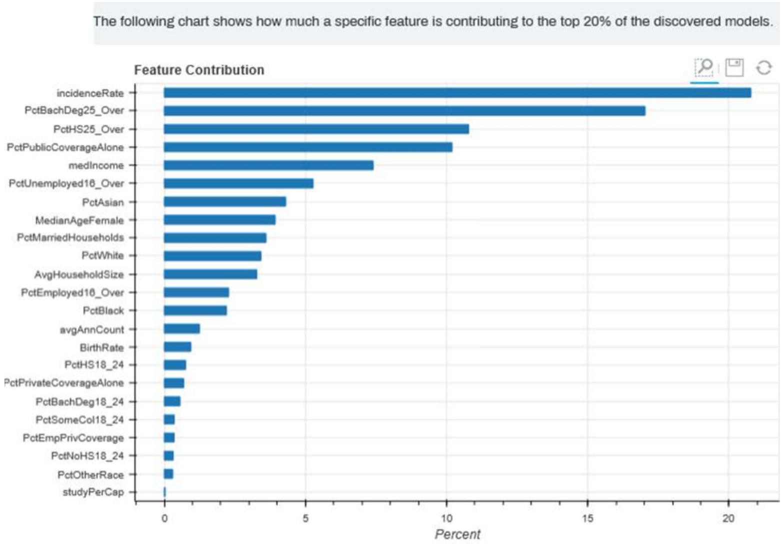


Metric	Scaled	Value
Num. points	-	609.00
Score	-	-15.655
Mean Abs Error	0.689	15.655
Median Abs Error	0.708	11.725
Root Mean Square Error	0.674	21.191
Coef Determination	0.470	0.470

Feature Importance graph for this model (top features)



Feature Contribution to the top 20% of the CausaLens discovered models



The following chart shows how much a specific feature is contributing to the top 20% of the discovered models.

6. Conclusion

Comparing the metrics of models in scope, it can be concluded that the CausaLens model ensemble.XGBRegressor with the given set of parameters provides best predictions, R-Squared metric of 0.470. This metric is found to be better than the Random Forest Regression, the best of four prediction models considered in this project.

The feature contribution to the top 20% of the CausaLens discovered models shows that IncidenceRate and PctBachDeg25_Over are the two most influential features – the same result that came from the Random Forest Regression model, therefore confirming the importance of these features for predicting the target DeathRate feature.

Please refer to file “Capstone Project CausaLens - Ana Potje - APPENDIX.pdf” for data dictionary details.