4. Predictive analytics and Conclusions

Predicting Heart Disease Mortality with Machine Learning Techniques
Lorenzo Negri, July 2018

Feature selection and engineering

The exploratory analysis and some experiments, lead to the following approach:

- Convert to indicator values (binary values) the *area__rucc* and *area__urban_influence* before feature selection, so that columns that contain categorical values will transform into a series of binary indicator columns that can be more easily used as single feature to be selected.
- Measure permutation feature importance to select features until obtaining the best combination of variables to feed the prediction model performance.
- Normalize data with ZScore transformation for *health__air_pollution_particulate_matter* , *health__homicides_per_100k* , *health__motor_vehicle_crash_deaths_per_100k* , *health__pop_per_dentist* , *health__pop_per_primary_care_physician* .
- Clip peaks with a threshold of 90% for upper values of all columns excluded *area_rucc_urban_popolation* (all types), *area_urban_micropolitan* (all types), area*_urban_influence_noncore* (all types). Clip peakes with a constant threshold of 0.22 for upper values of *demo__pct_aged_65_years_and_older*.

Managing Missing data

Based on the analysis and experiments done, the best solution obtained was using different techniques of filling missing data.

The mixed approach: lower number of missing data columns uses one type of managing missing data and higher, another type. Only *health__homicides_per_100k* uses a different type for its own:

- Probabilistic PCA: for features with missing data <= 6%
- MICE: for features with missing data > 6%
- Mean of data: for *health__homicides_per_100k* feature missing data

Probabilistic PCA: Replaced the missing values by using a linear model that analyses the correlations between the columns and estimates a low-dimensional approximation of the data, from which the full reconstructed data. The underlying dimensionality reduction is a probabilistic form of Principal Component Analysis (PCA).

MICE: For each missing value, it assigns a new value, calculated by using a method described in the statistical literature as "Multivariate Imputation using Chained Equations" or "Multiple Imputation by Chained Equations". With a multiple imputation method, each variable with missing data conditionally

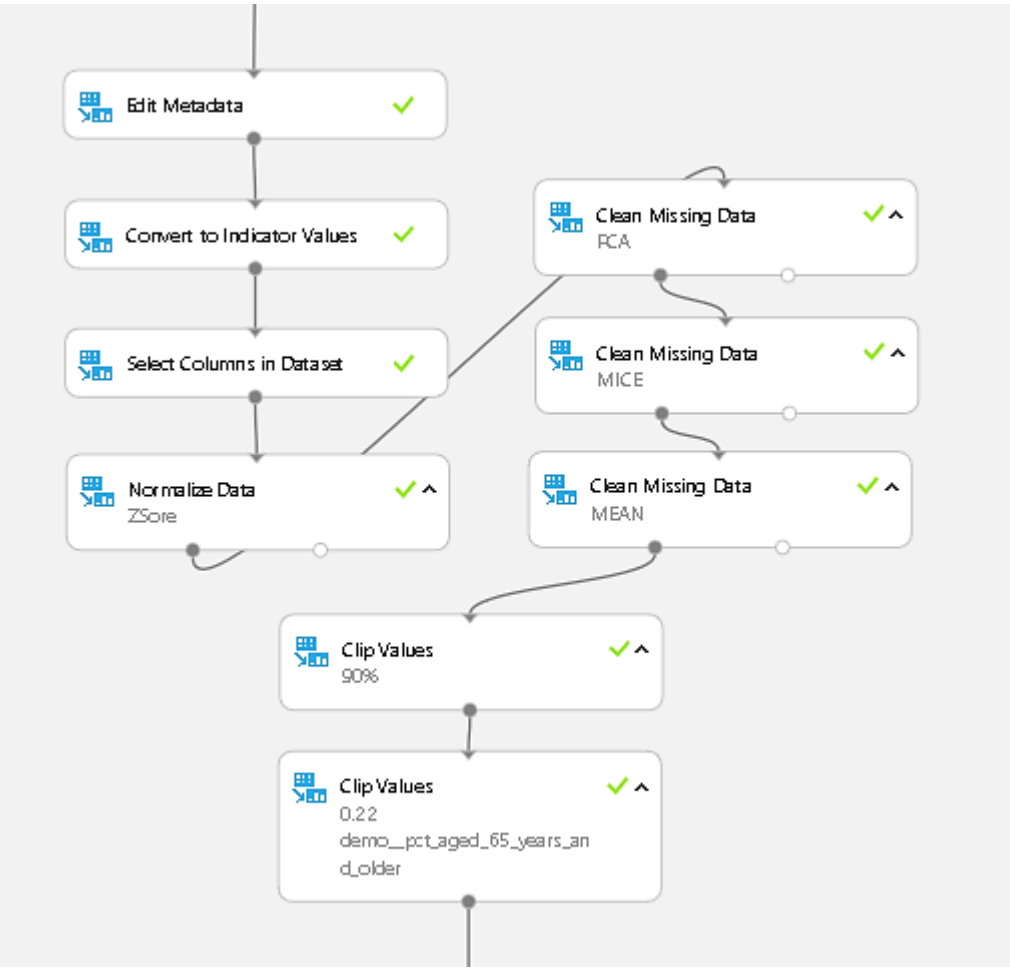
modelled using the other variables in the data, before filling in the missing values. In contrast, in a single imputation method (such as replacing a missing value with a column mean) as made in a single pass over the data to determine the fill value.

Mean: Replace all missing data with calculated mean of the column.

The following are the Root Mean Square Error (RMSE) for different techniques used to achieve the results of the final ML model obtained with the *test_values* dataset and scored with client future data:

Type	Result
All MICE	33.6952
All Probabilistic	33.9396
Mixed approach	**33.5059**

Azure Machine Learning data streamflow for preparing features for regression:



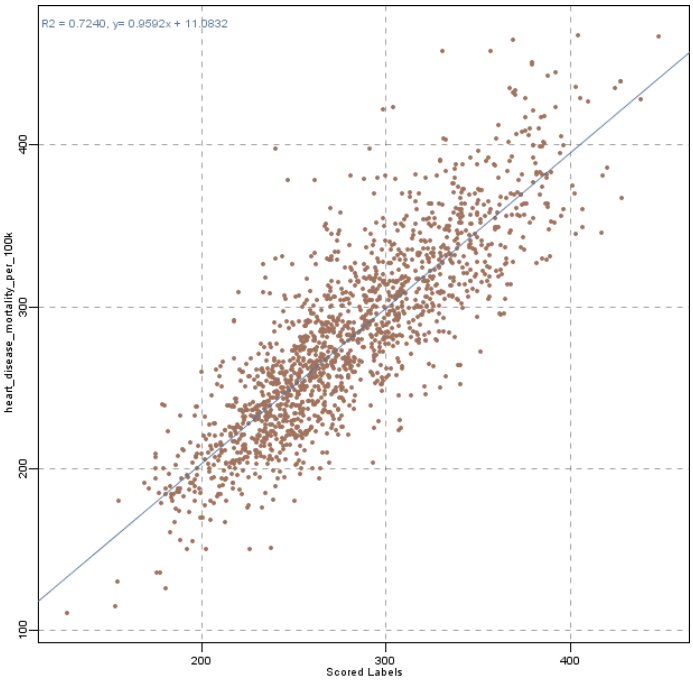
Overfitting Test Data

To avoid overfitting test data and getting worst result when scoring future data, the following showed to be efficient when tested with cross validation:

- Label values (*heart_disease_mortality_per_100k*) joined to the train dataset after replacing missing values.
- Binning, clipping, and outlier removal have to be selected carefully to avoid overfitting.
- Selection of features of the converted categorical to indicator values (binary) needs more experiments to tune the best-mixed solution.

Regression

After managing the missing data and selected the features in the optimal way measured by performance feature importance, creating some regressions machine learning models is the best way to find the right one to use. Based on the score results of Root Mean Square Error (RMSE), a Boosted Decision Tree Regression model is chosen to predict the *heart_disease_mortality_per_100k*, from which will be generated the integer predicted rate to score.



The model trained with 50% of the data, and tested with the remaining 50% generated a scatter plot showing the predicted *heart_disease_mortality_per_100k* and the actual rates.

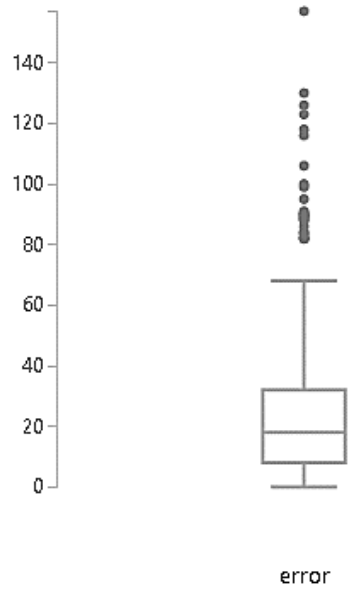
This plot shows a clear linear relationship between predicted and actual values in the test dataset. The metrics for the test results are:

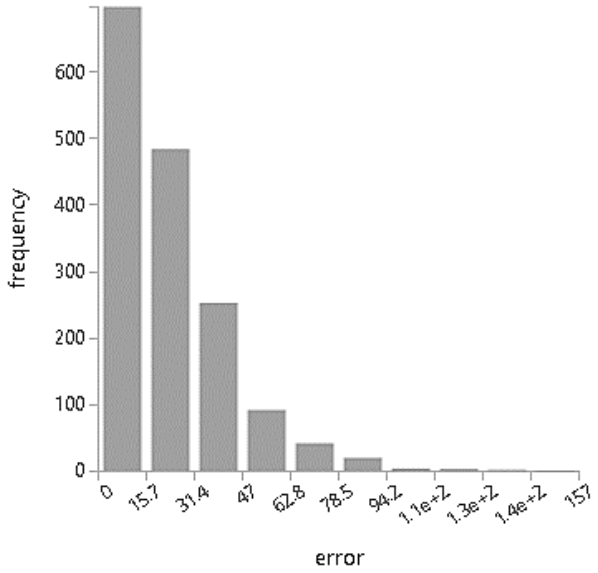
Mean Absolute Error	22.917346
Root Mean Squared Error	30.201038
Relative Absolute Error	0.494744
Relative Squared Error	0.277389
Coefficient of Determination	0.724010

The model predicts sufficiently well, and in a stable way with less overfitting effect, the metrics in the cross validation model for Root Mean Squared Error.

Root Mean Squared Error stats:

Mean	28.4394
Median	28.2883
Max	30.4298
Standard Deviation	1.565734





Conclusion

This analysis has shown that the **target variable** (*heart_disease_mortality_per_100k*) of a dataset consisting of features of United States county-level area info, demographic and socioeconomic and from its characteristics can be predicted. The model can be tuned further to obtain better results.

The model shows best results when there is less missing data. The main features that are critical for the algorithm to obtain better results are *pct_physical_inacticity*, *pct_adults_bachelors_or_higher*, *pct_adult_smoking*, *pct_diabetes*. Secondary features, such as categorical ones can help further when converted to indicator values and nominal value are selected optimally. Other features are important when appropriately managed and engineered as *pct_aged_65_years_and_older*. Further selection of features is important to have best results.