Enhancing Public Health Outcomes: Developing a Cancer & Stroke Prediction Model.

Rohan Chintalapati

Department of Mechanical and Industrial Engineering, University of Massachusetts, Amherst

MIE - 622: Predictive Analytics & Statistical Learning

Professor Michael Prokle

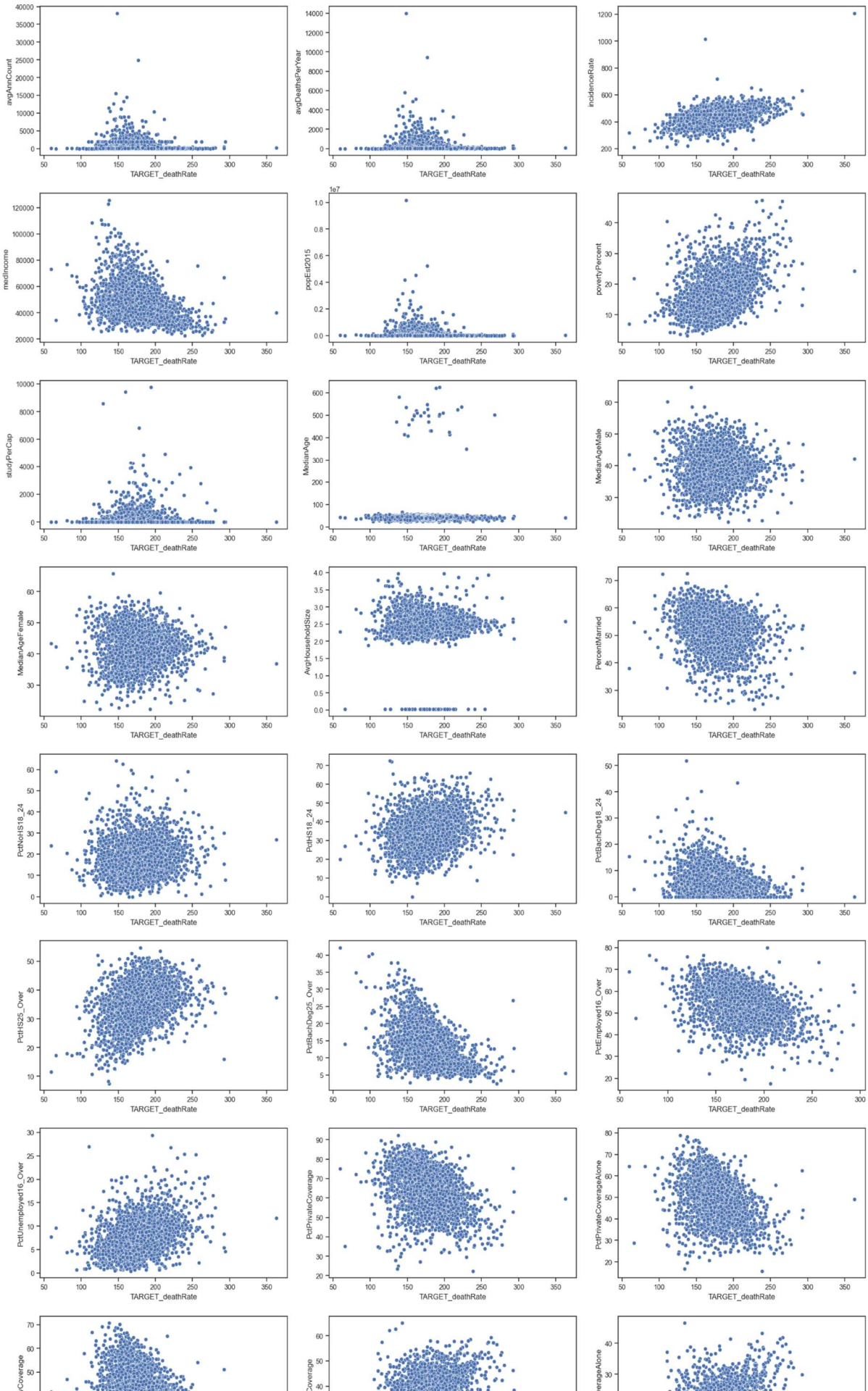
Cancer Prediction Model

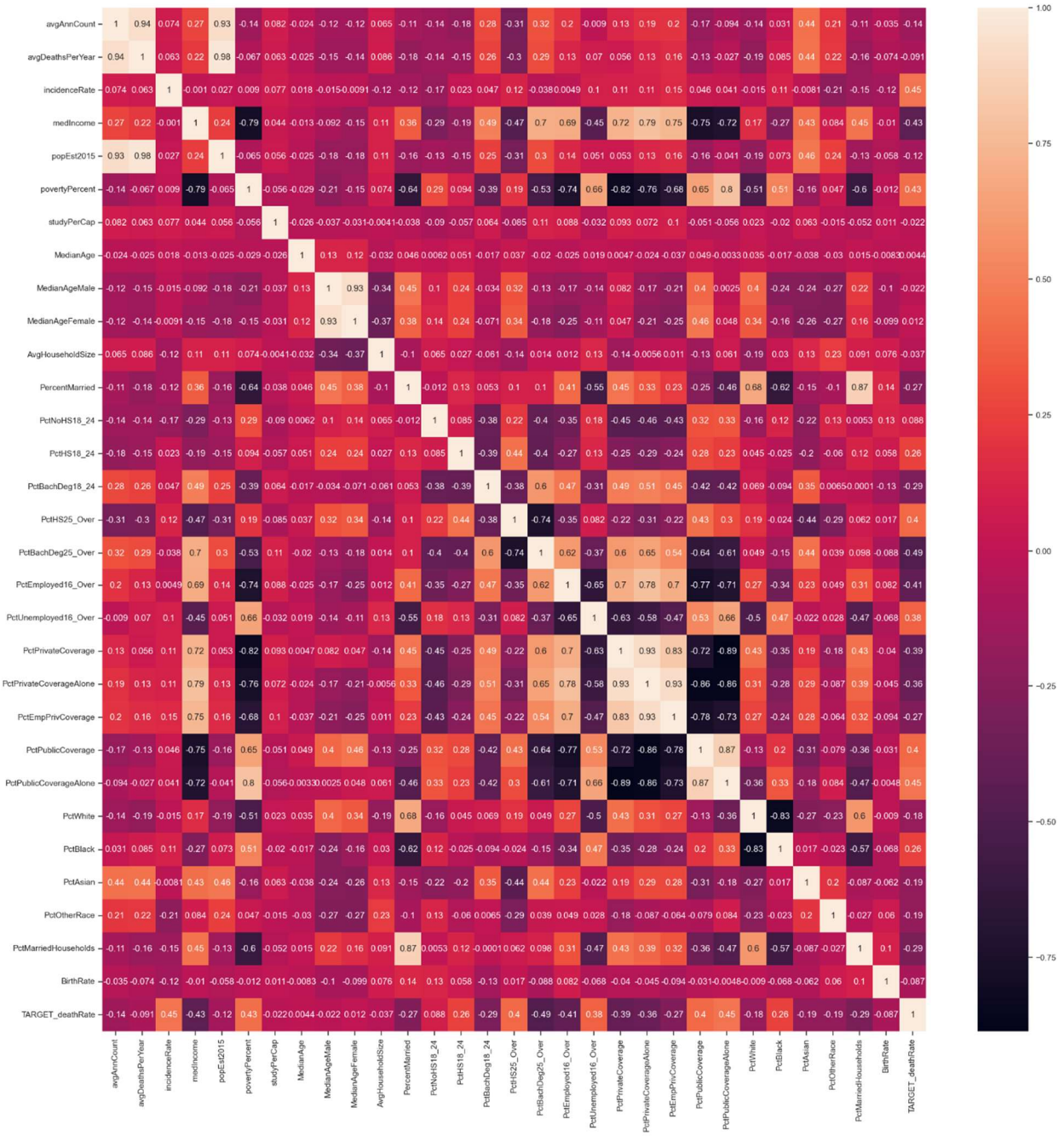
Executive Summary

The analysis was performed on a cancer dataset to identify factors that contribute to the target variable, "TARGET_deathRate." The dataset contained 3047 rows and 34 columns, with various features such as demographic, socioeconomic, and health-related attributes.

Key Insights:

Correlation Analysis: Several features were found to be positively correlated with the target variable, such as incident rate, povertyPercent, PctHS25_over, PctHS18_24, PctUnemployed16_Over, PctPublicCoverage, and PctPublicCoverageAlone. On the other hand, medianIncome, PercentMarried, PctEmpPrivCoverage, PctPrivateCoverage, PctPrivateCoverageAlone, PctBatchDeg25_Over, PctEmplyed16_Over, and pctMarriedHousehold were negatively correlated with the target variable. We obtained this data from the scatterplot and heatmap below (Data Exploration).


avgAnnCount
avgDeathsPerYear
incidenceRate
medIncome
popEst2015
povertyPercent
studyPerCap
MedianAge
MedianAgeMale
MedianAgeFemale
AvgHouseholdSize
PercentMarried
PctNoHS18_24
PctHS18_24
PctBachDeg18_24
PctHS25_Over
PctBachDeg25_Over
PctEmployed16_Over
PctUnemployed16_Over
PctPrivateCoverage
PctPrivateCoverageAlone
TARGET_deathRate



Outliers Handling: Some features, including avgAnnCount, avgDeathPerYear, IncidentRate, MedianAge, and studyPerCap, had outliers. To mitigate the impact of outliers, the upper whisker value was used for replacement.

Null Value Handling: Null values were found in PctPrivateCoverageAlone and PctEmployed16_Over columns. The missing values were filled using the mean values of the respective columns.

Feature Selection: Stepwise selection was performed to identify the best set of variables that significantly contribute to the target variable.

The selected features were:

```
Best set of variables:
['const', 'PctBachDeg25_Over', 'incidenceRate', 'povertyPercent', 'PctHS18
_24', 'PctOtherRace', 'avgDeathsPerYear', 'popEst2015', 'MedianAgeFemale',
'avgAnnCount', 'PctBlack', 'BirthRate', 'PctPrivateCoverage', 'PctMarriedH
ouseholds', 'PercentMarried', 'PctEmployed16_Over', 'PctEmpPrivCoverage',
'PctHS25_Over', 'studyPerCap', 'MedianAgeMale', 'PctAsian']
```

Linear Regression Model: A linear regression model was built using the selected features. The model's performance was evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2). The R-squared value indicated that the model explained approximately 54.4% of the variance in the target variable.

Alternative Model - **Random Forest Regression**: A Random Forest Regressor was also trained on the data. The performance of this model was compared with the linear regression model. The Random Forest Regressor outperformed the linear regression model, achieving an R-squared value of approximately 57.8%.

Results

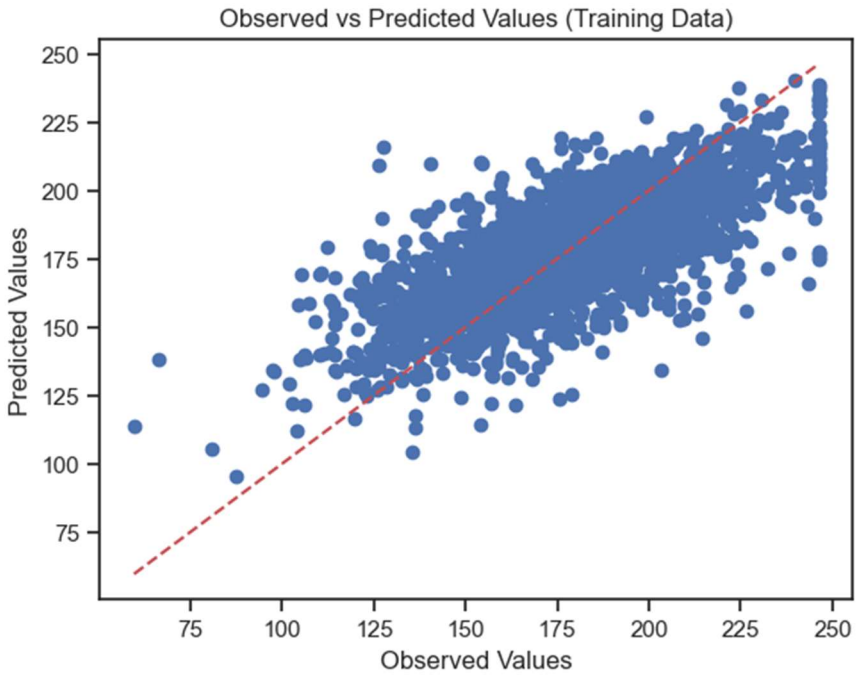
The evaluation metrics calculated on the test set are as follows:

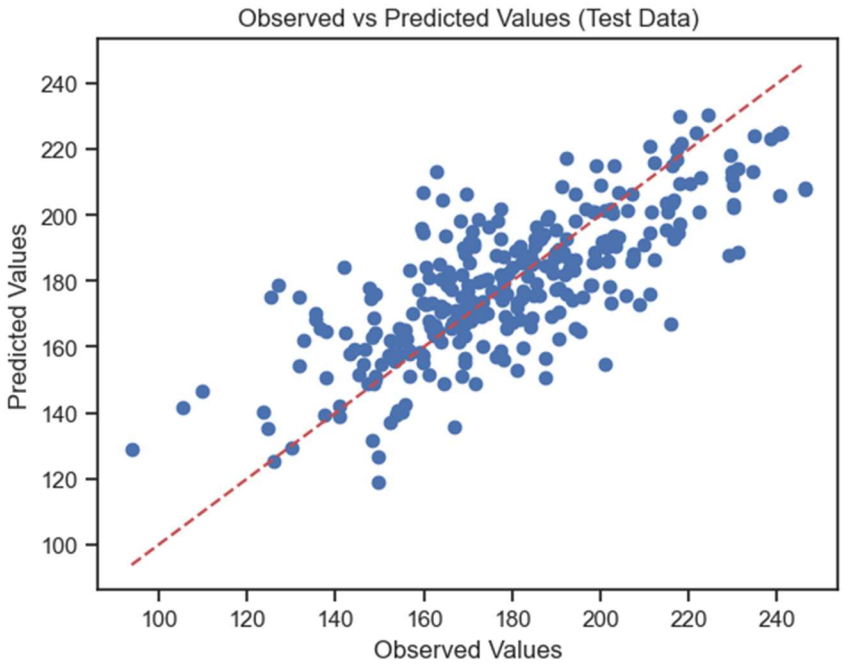
Mean Squared Error (MSE): 315.10

Mean Absolute Error (MAE): 13.57

Root Mean Squared Error (RMSE): 17.75

R-squared (R2): 0.57





Random Forest Regressor Model:

The Random Forest Regressor is an ensemble learning method that uses multiple decision trees to make predictions. It often performs well in capturing complex relationships between features and the target variable.

Evaluation Metrics: The evaluation metrics calculated on the test set are as follows:

Mean Squared Error (MSE): 306.31

Mean Absolute Error (MAE): 13.34

Root Mean Squared Error (RMSE): 17.50

R-squared (R2): 0.58

Comparison:

The Random Forest Regressor outperformed the OLS Regression model in terms of the evaluation metrics. It achieved a slightly lower MSE, MAE, and RMSE and a higher R-squared value, indicating that it explained more variance in the target variable compared to the OLS model.

Overall, based on the evaluation metrics, the Random Forest Regressor is a better model for predicting the "TARGET_deathRate" compared to the OLS Regression model. However, to make a more robust conclusion, further analysis, such as cross-validation and additional model tuning, may be necessary.

Conclusion

The analysis revealed that several demographic, socioeconomic, and health-related features influence the target variable, "TARGET_deathRate." The Random Forest Regressor provided better predictions compared to the linear regression model. The insights from this analysis could be valuable in understanding the factors contributing to cancer death rates and aid in the development of targeted interventions to improve public health outcomes. However, further exploration and validation with additional data could be helpful to enhance the model's accuracy and generalizability.

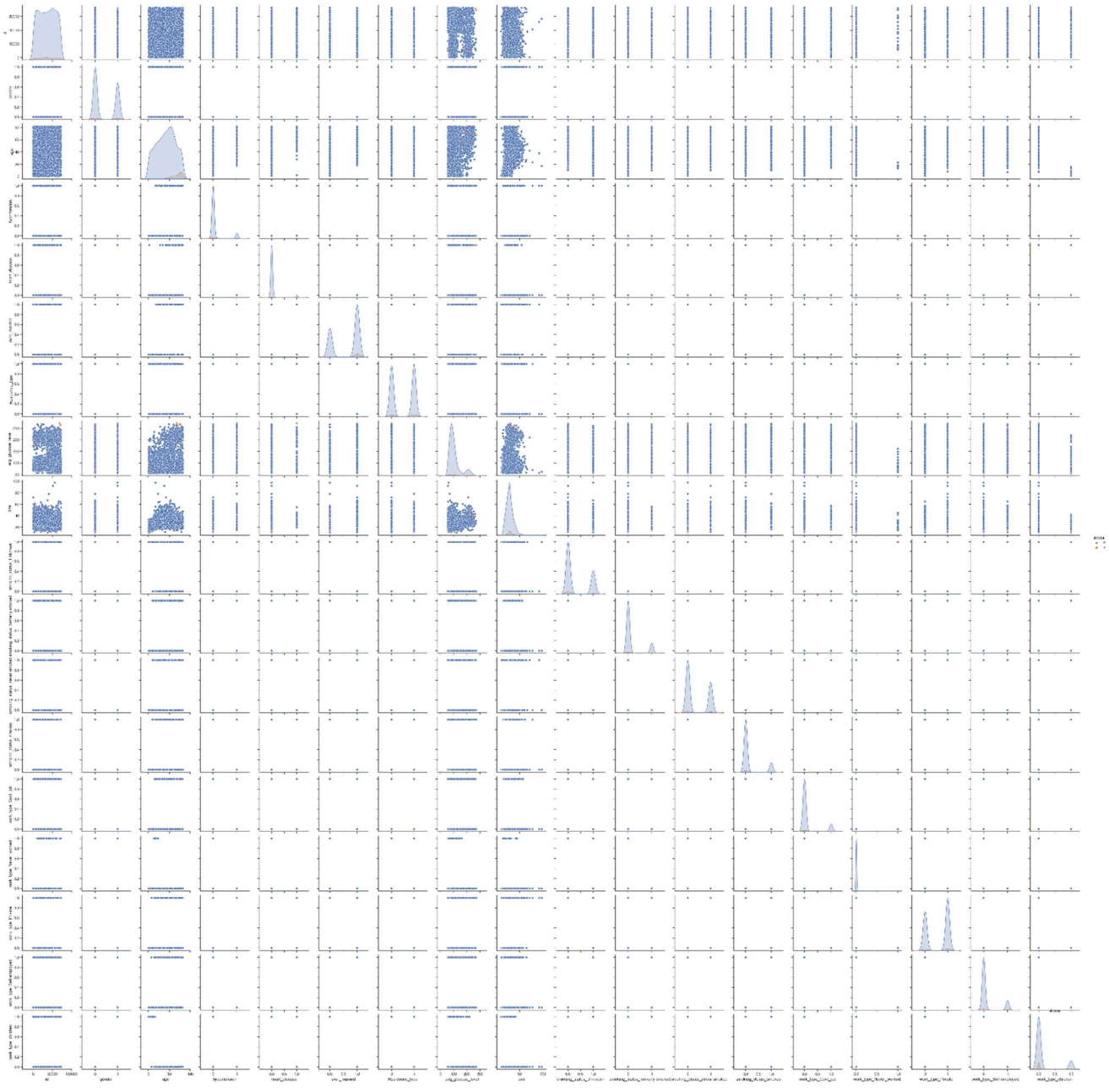
Stroke Prediction Model

Executive Summary

The goal of this project was to create a stroke prediction model using the provided dataset. The dataset contains information about various features such as age, gender, hypertension, heart disease, average glucose level, BMI, smoking status, work type, and residence type. The target variable is whether the individual had a stroke or not.

Two classification models were trained and evaluated on the dataset: Logistic Regression and Random Forest Classifier. Both models were used to predict the likelihood of stroke based on the given features. The Random Forest model achieved a higher accuracy score compared to the Logistic Regression model.

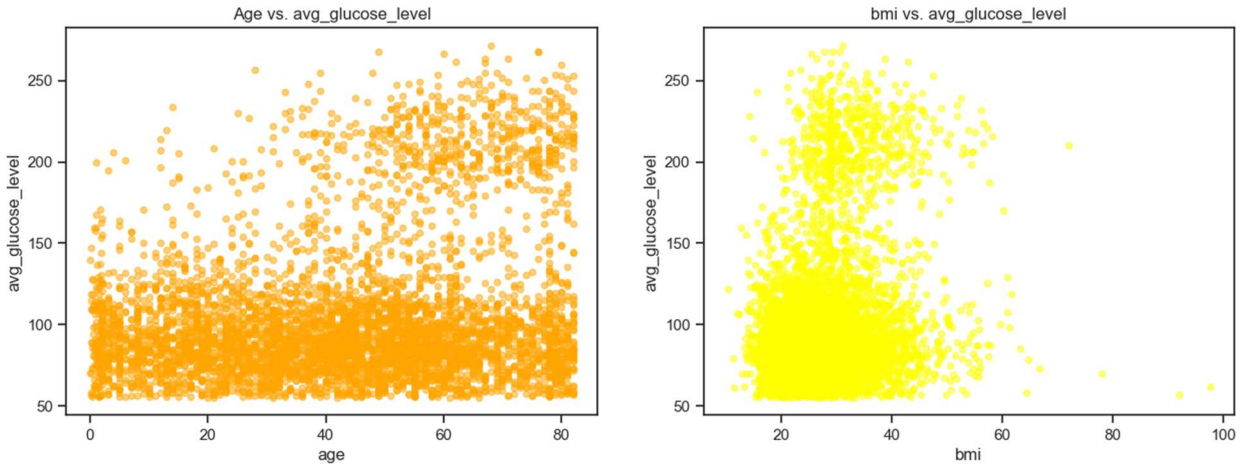
Data Exploration



Most of the features follow the normal distribution while the PctWhite feature is left skewed. PctBlack, PctAsian, PctOtherRace, AvgAnnount, avgDeathsPerYear, popEst2015, studyPerCap, MedianAge features are right skewed according to the above scatterplot.

The scatter plot titled "Age vs. avg_glucose_level" shows the distribution of data points with 'age' on the x-axis and 'avg_glucose_level' on the y-axis. The orange dots represent data points in this plot. The scatter plot suggests a relatively uniform distribution of data points without a clear linear relationship between age and average glucose level.

The scatter plot titled "bmi vs. avg_glucose_level" shows the distribution of data points with 'bmi' on the x-axis and 'avg_glucose_level' on the y-axis. The yellow dots represent data points in this plot. Similar to the first plot, this scatter plot also shows a relatively uniform distribution of data points without a clear linear relationship between BMI and average glucose level.



Understanding and Interpretation of Results:

Logistic Regression Results

Testing Score: The logistic regression model achieved an accuracy of approximately 94.5% on the test dataset. However, accuracy alone may not be sufficient to evaluate the model's performance, especially when dealing with imbalanced classes.

Precision and Recall: The precision for class 1 (stroke occurrences) is reported as 0.00, indicating that the model did not correctly predict any true positive cases. The recall for class 1 is also 0.00, which means the model failed to identify any actual stroke cases correctly. These low values suggest that the model struggled to classify positive cases correctly, likely due to the class imbalance.

F1-score: The F1-score for class 1 is 0.00, which is the harmonic mean of precision and recall. Since both precision and recall are 0.00, the F1-score for class 1 is also 0.00, highlighting the model's difficulty in correctly predicting positive cases.

Confusion Matrix: The confusion matrix shows that the model correctly classified all negative cases (no stroke occurrences) but failed to correctly predict any positive cases (stroke occurrences).

Random Forest Classifier Results:

Training Score: The Random Forest model achieved a near-perfect accuracy of approximately 99.98% on the training dataset. While this high score indicates good performance on the training data, it raises concerns about potential overfitting due to the class imbalance.

Testing Score: The accuracy on the test dataset dropped to around 93.3%, which is still decent, but not as high as the training accuracy. This discrepancy between training and testing scores also indicates the presence of overfitting.

Precision and Recall: Like the Logistic Regression model, the Random Forest model struggles to predict positive cases. The precision for class 1 is low, indicating that out of the positive predictions made by the model, only a small fraction were true positive cases. The recall for class

1 is also low, suggesting that the model did not identify a significant number of actual stroke cases.

F1-score: The F1-score for class 1 is close to 0.00, indicating that the model's ability to correctly predict positive cases is very poor.

Confusion Matrix: The confusion matrix shows that the model correctly classified the majority of negative cases but struggled to predict positive cases, resulting in a relatively large number of false negatives (actual stroke cases misclassified as non-stroke cases).

Interpretation

Both models exhibited challenges in predicting positive cases (stroke occurrences) due to the class imbalance in the dataset. The imbalance led the models to heavily favor the majority class (non-stroke occurrences) while neglecting the minority class (stroke occurrences).

Quantitative Insights

1. The Logistic Regression model failed to predict any true positive cases, leading to 0.00 precision, recall, and F1-score for class 1.

2. The Random Forest model, despite its high training accuracy, suffered from overfitting and struggled to predict positive cases, resulting in low precision, recall, and F1-score for class 1.

Possible Improvements

1. Class Imbalance Handling: Addressing the class imbalance issue through techniques like oversampling, undersampling, or using SMOTE could improve the models' performance in predicting positive cases.

2. Feature Engineering: Exploring additional features or transforming existing ones might capture more relevant information and improve model performance.

3. Hyperparameter Tuning: Fine-tuning the model hyperparameters using techniques like Grid Search or Random Search could optimize performance.

Practical Relevance

Given the models' current limitations, they may not be suitable for making critical medical decisions. However, as screening tools, they can still aid healthcare professionals in identifying individuals at higher risk of stroke. For practical use, further refinement of the models to enhance their ability to predict positive cases is essential.

If your company can only start with one the two use cases (cancer or stroke), which would you recommend to be pursued first based on your analysis and the preliminary data given?

Based on the analysis and preliminary data given, I would recommend pursuing the Cancer Prediction Model as the first priority for your company. Here are the reasons for this recommendation:

1. Dataset Size and Features: The cancer dataset contains 3047 rows and 34 columns with various demographic, socioeconomic, and health-related attributes. This larger dataset provides more data points for training and testing the predictive model, which can lead to better generalization and robustness.

2. Strong Predictive Power: The Random Forest Regressor model achieved an R-squared value of approximately 57.8% on the test set, indicating that it explained a significant portion of the variance in the target variable ("TARGET_deathRate"). This suggests that the model has reasonable predictive power for cancer death rates.

3. Insights for Public Health: A well-developed cancer prediction model can provide valuable insights into factors influencing cancer death rates. Understanding these factors can help identify regions or populations with higher risks and guide the development of targeted interventions to improve public health outcomes.

4. Potential Impact: Cancer is a significant public health concern, and predicting cancer death rates can have a substantial impact on healthcare planning, resource allocation, and policy decisions. Accurate predictions can aid in focusing resources on high-risk populations and implementing preventive measures.

5. Competitive Advantage: Developing a successful cancer prediction model can give your company a competitive advantage in the healthcare industry. Accurate predictions and valuable insights can attract collaborations with healthcare providers, researchers, and policymakers.

While the stroke prediction model is also essential, the analysis showed some limitations in accurately predicting positive cases due to class imbalance and overfitting. Addressing these issues and refining the stroke prediction model would require further effort and additional data. Thus, starting with the cancer prediction model would be a prudent choice based on the current state of the models and data. Once the cancer prediction model is well-established and effective,

your company can then shift its focus and resources to further improving the stroke prediction model.

References

1. https://towardsdatascience.com/understanding-regression-metrics-and-its-business-implementation-c4e8a32bb74e

2. "Introduction to Statistical Learning" by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani