Heart Disease Prediction

Group 8 - Juzheng Shi, Sneha Batchu, Sony Kumari, Udisha Madnani

Introduction

Heart disease is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy.

In the United states alone, heart diseases claims roughly around 647000 lives each year and thus is regarded as the leading cause of death.

There are different symptoms of coronary heart diseases, and the majority of people only learn about this disease once they suffer from symptoms like chest pain, a heart attack or sudden cardiac arrest.

Business Problem Description

The Centers for disease control and Prevention has identified high blood pressure, high blood cholesterol, and smoking as three key risk factors for heart disease. Cardiac disease affects not just people's health but economies and costs of countries as well. The cardiac disease accounts for a third of all healths in people over 35.

People with cardiac disease have more days of unplanned absence from work and are less productive while at work compared to the general population. The researchers estimated the economic impact of stopping future cases of coronary heart disease over the next 10 years (2020-2029). Preventing all future cases would nearly save USD $15 billion in GDP which is equivalent to almost 51000 USD for every case avoided.

Objective

We intend to predict heart disease or attacks, using the results of a survey for preventative health screening of heart diseases, and using the following techniques:

- Data Cleaning
- Data Analysis
- Feature Engineering
- Preprocessing
- Classification models : Logistic Regression, Random Forest, LightGBM
- Final Evaluation

Data Source

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. This original dataset contains responses from 441,455 individuals and has 330 features. These features are either questions directly asked of participants or calculated variables based on individual participant responses. The refined dataset contains 253680 rows and 22 features.

Link to the data source - https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset

Data Description

- **Categorical Features:**
  - *HighBP* Whether the individual has High BP or not
  - *HighChol* Whether the individual has High Cholesterol or not
  - *CholCheck* Whether the individual has got Cholesterol levels checked or not
  - *Smoke* Whether the individual smokes or not
  - *Stroke* Whether the individual has ever had a heart stroke or not
  - *Diabetes* Whether the individual has diabetes or not
  - *PhysActivity* Whether the individual is physically active or not
  - *Fruits* Whether the individual consumes fruits or not
  - *Veggies* Whether the individual consumes vegetables or not
  - *Alcohol* Whether the individual consumes alcohol or not
  - *Healthcare* Whether the individual has any healthcare plan or not
  - *DocCost* Whether the individual has incurred any expenses for doctor visit or not
  - *DiffWalk* Whether the individual faces any difficulty in walking or not

- **Continuous Features:**
  - *BMI* BMI level of the individual

- **Ordinal Features:**
  - *Age* Age of the individual
  - *Sex* Sex of the individual
  - *Education* Educational qualification of the individual
  - *Income* Income level of the individual
  - *GenHlth* Rating of the individual's General Health
  - *MentHlth* Rating of the individual's Mental Health
  - *PhysHlth* Rating of the individual's Physical Health

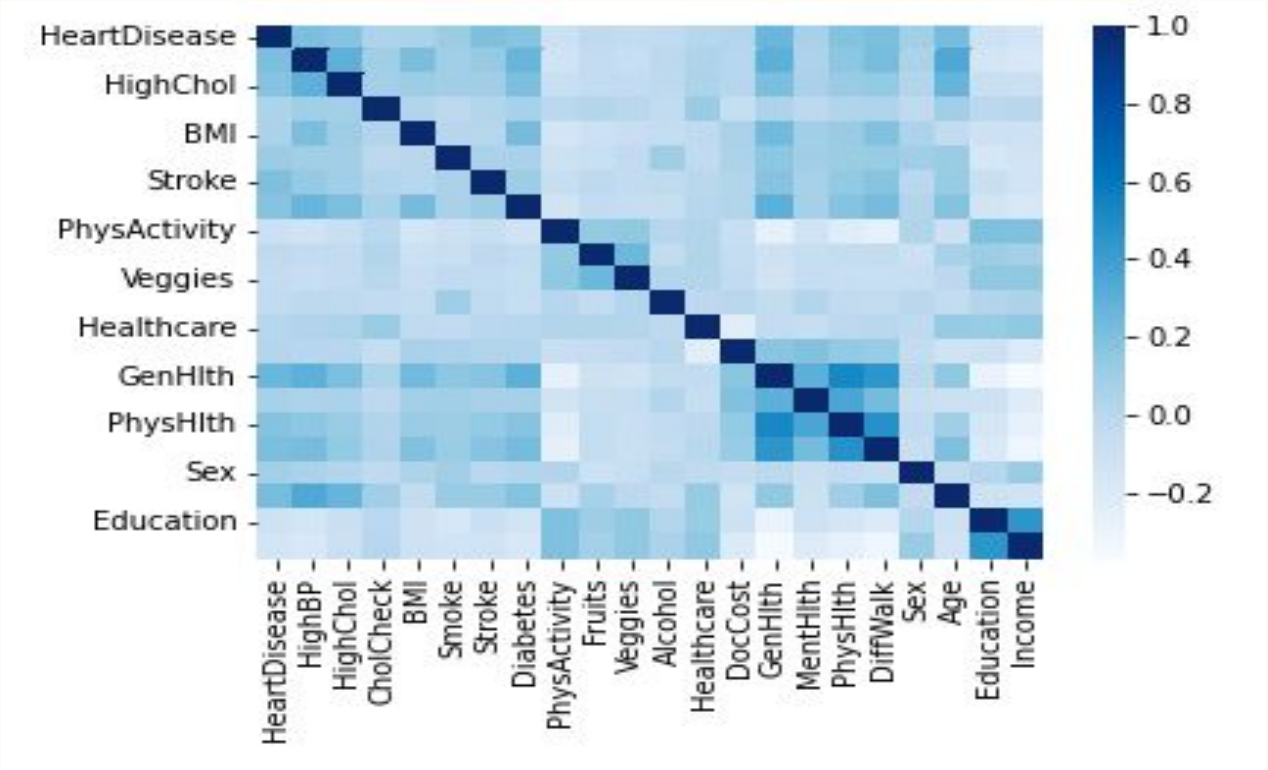
Data Description

- Ordinal and continuous features

```
BMI Length: 84
BMI : [12. 13. 14. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29.
 30. 31. 32. 33. 34. 35. 36. 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47.
 48. 49. 50. 51. 52. 53. 54. 55. 56. 57. 58. 59. 60. 61. 62. 63. 64. 65.
 66. 67. 68. 69. 70. 71. 72. 73. 74. 75. 76. 77. 78. 79. 80. 81. 82. 83.
 84. 85. 86. 87. 88. 89. 90. 91. 92. 95. 96. 98.]
Diabetes Length: 3
Diabetes : [0. 1. 2.]
GenHlth Length: 5
GenHlth : [1. 2. 3. 4. 5.]
MentHlth Length: 31
MentHlth : [ 0.  1.  2.  3.  4.  5.  6.  7.  8.  9. 10. 11. 12. 13. 14. 15. 16. 17.
 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30.]
PhysHlth Length: 31
PhysHlth : [ 0.  1.  2.  3.  4.  5.  6.  7.  8.  9. 10. 11. 12. 13. 14. 15. 16. 17.
 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30.]
Age Length: 13
Age : [ 1.  2.  3.  4.  5.  6.  7.  8.  9. 10. 11. 12. 13.]
Education Length: 6
Education : [1. 2. 3. 4. 5. 6.]
Income Length: 8
Income : [1. 2. 3. 4. 5. 6. 7. 8.]
```

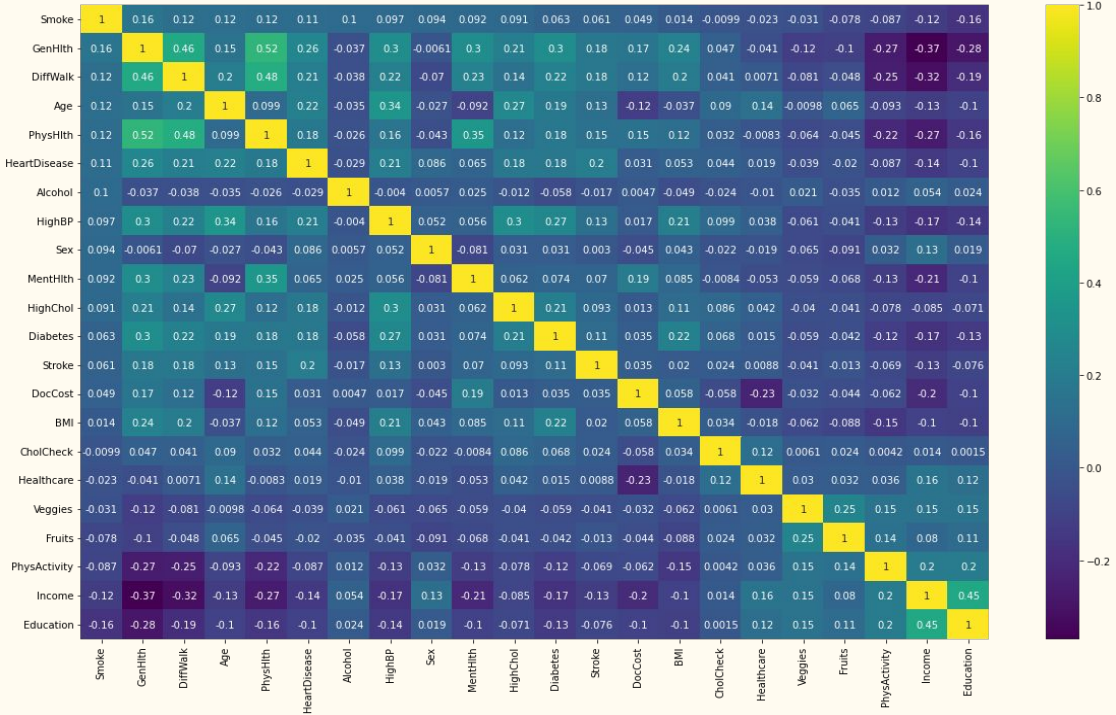
Exploratory Data Analysis

Checking correlation



Exploratory Data Analysis

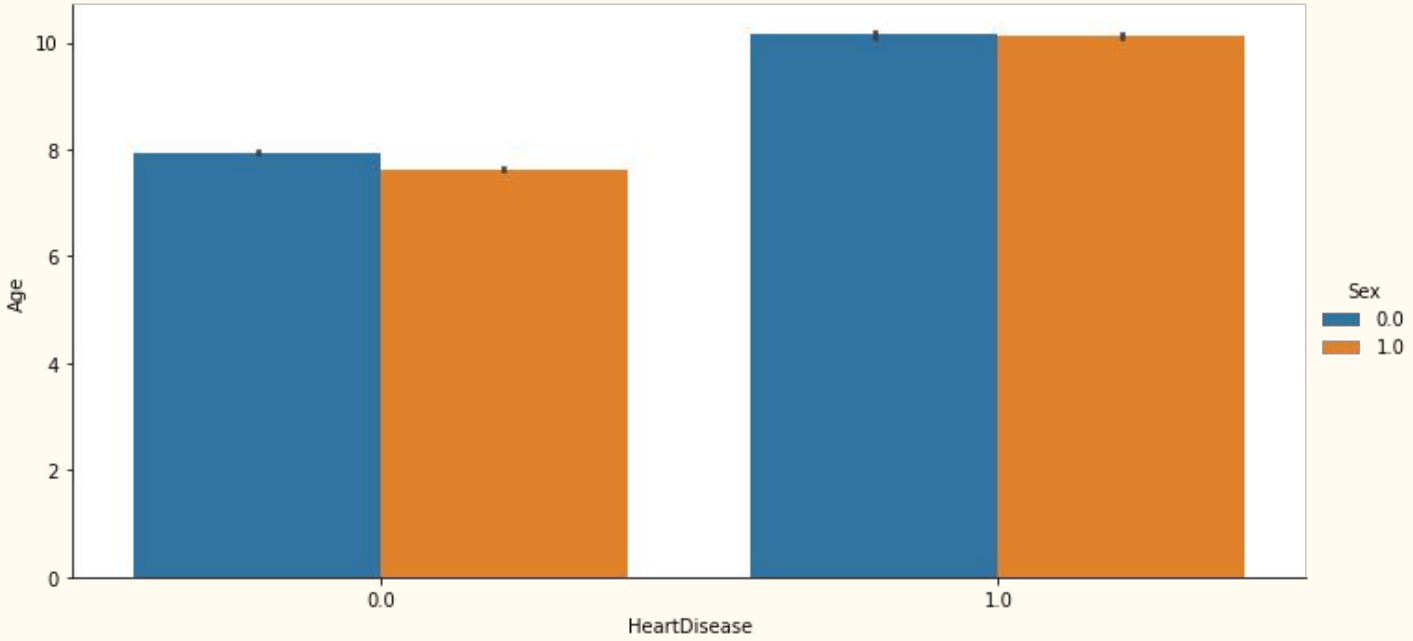
Quality correlation matrix



- Here we can infer that "PhysHlth" has strong positive correlation with "GenHlth" whereas it has strong negative correlation with "income".
- "HealthCare" and "DocCost" has almost no correlation with "HeartDisease".
- Since correlation is zero we can infer there is no linear relationship between these two predictors. However, it is safe to drop these features in case you're applying Linear Regression model to the dataset.

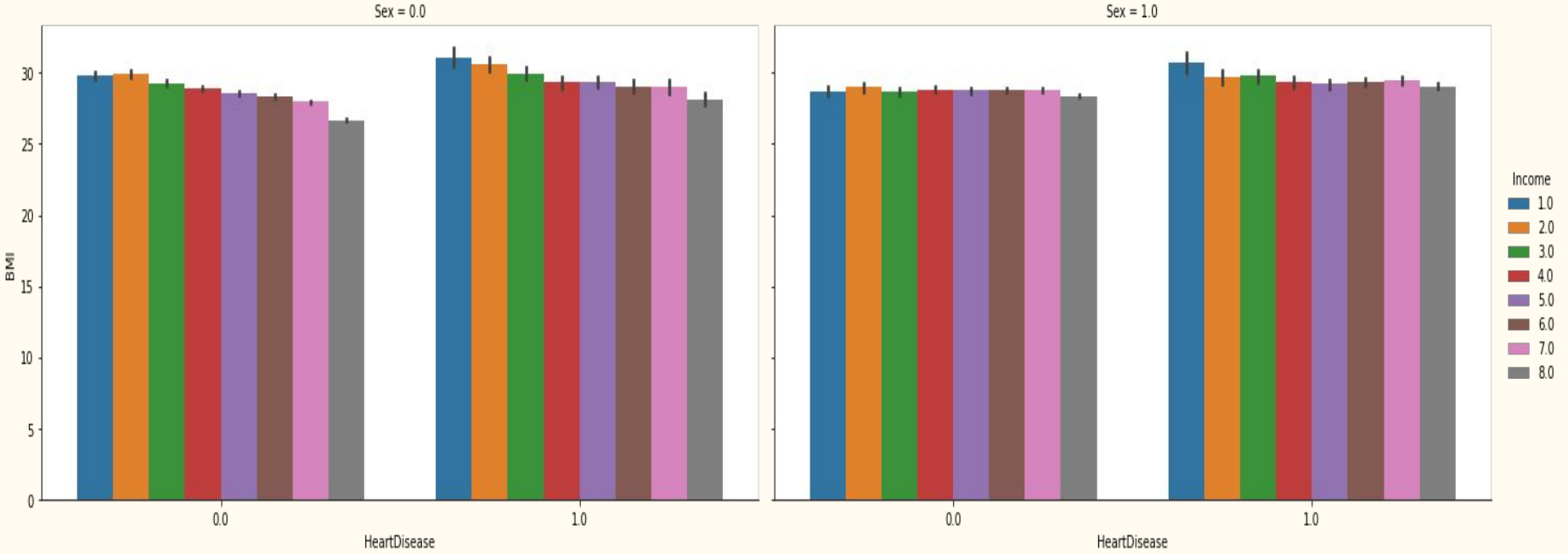
Exploratory Data Analysis

Heart disease with respect to age and sex



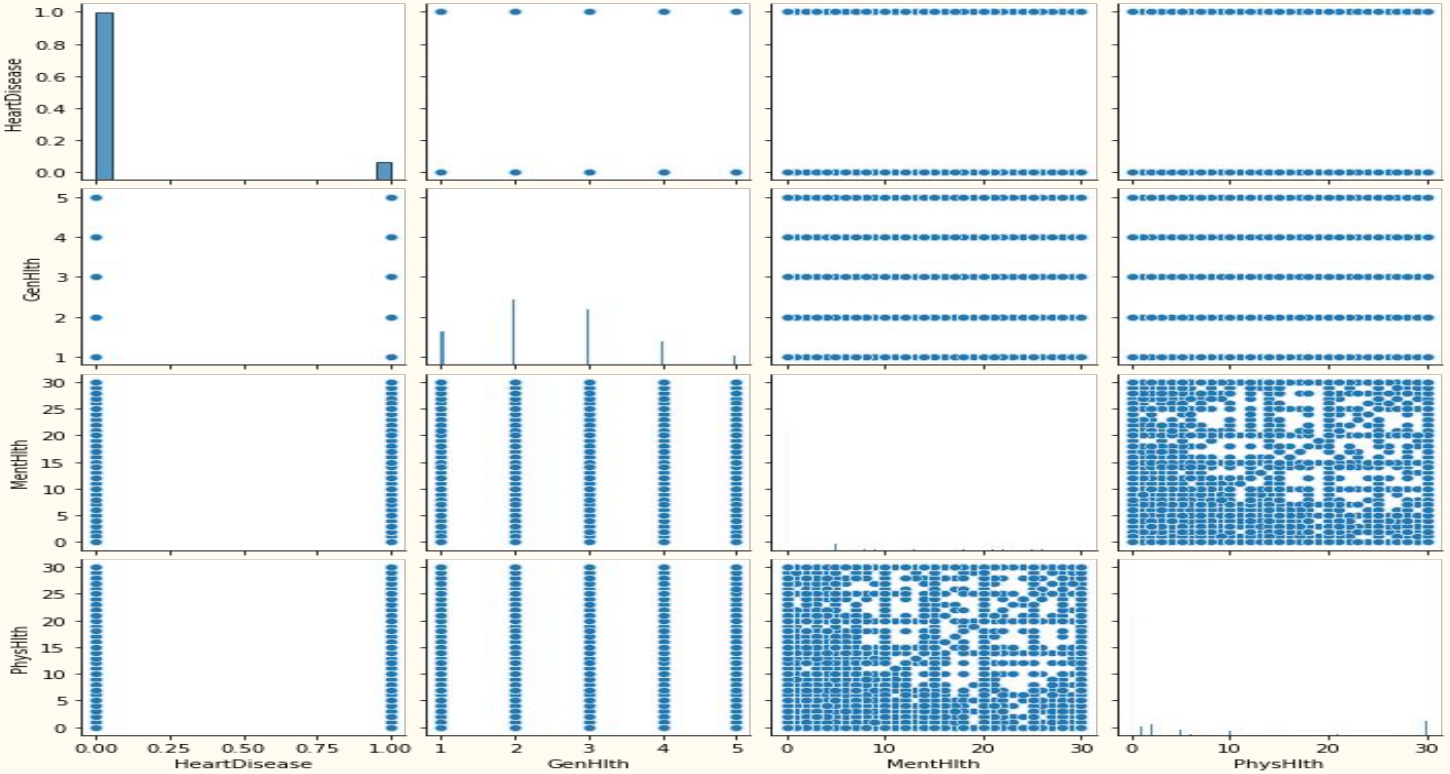
Exploratory Data Analysis

Heart disease with respect to BMI, Income and Sex

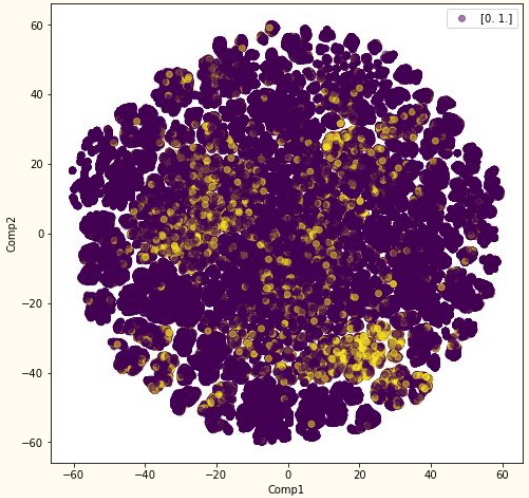


Exploratory Data Analysis

Pairplot of Heart disease, Mental health, Physical health and General health

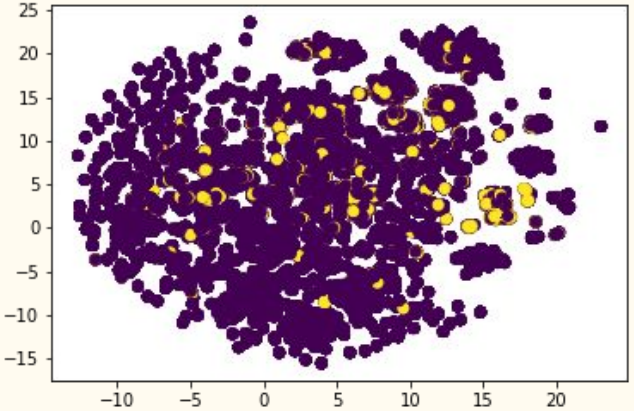


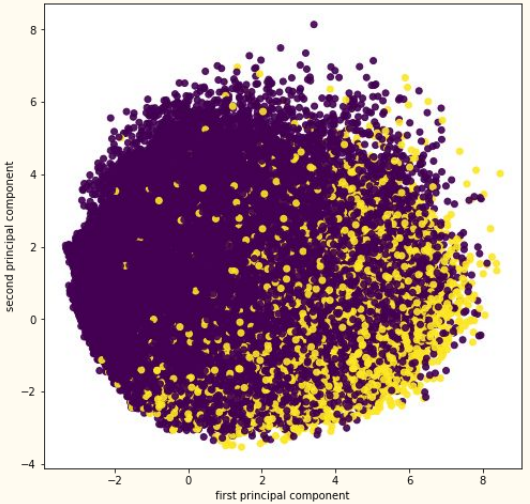
Dimensionality Reduction



T-distributed Stochastic Neighbor Embedding

Uniform Manifold Approximation & Projection

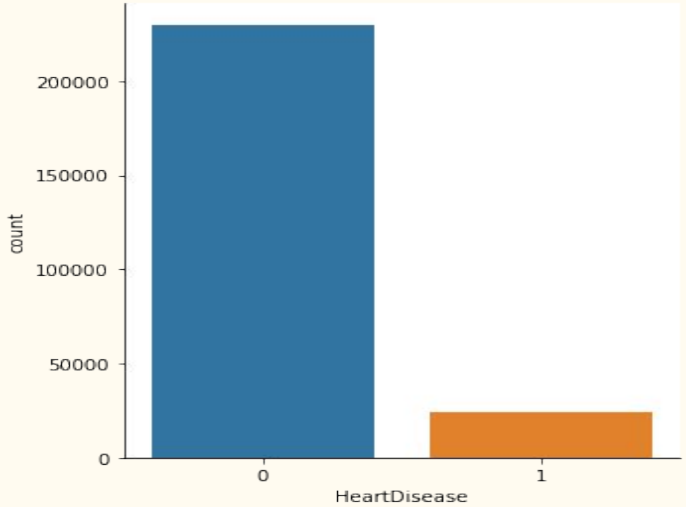




Principal component analysis

Preprocessing

- As we can see from the distribution of the labels, this data set is imbalanced.
- To combat the imbalanced classes in the data, we resampled the data set, deleted samples from the over-represented label(under-sampling).
- For all the classification models, we had run them on balanced data set and imbalanced data set.
- We chose accuracy and recall as our metrics for model evaluation.



Feature Engineering

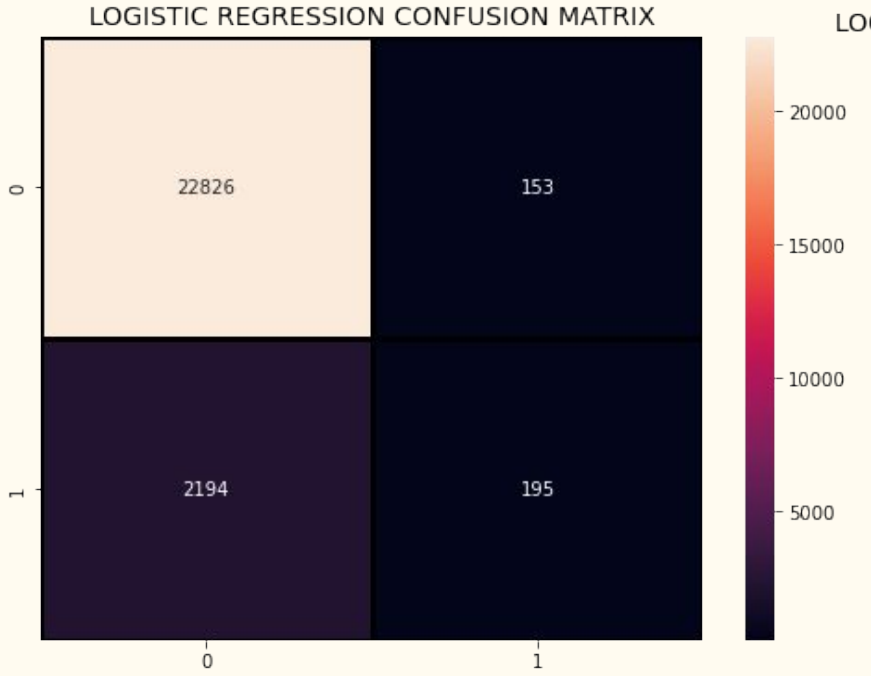
- Different feature engineering strategies
  a. Using one hot encoding on categorical features and standard scaler on numeric and ordinal features
  b. Using target encoding to replace categorical features with numbers and construct some interactive features

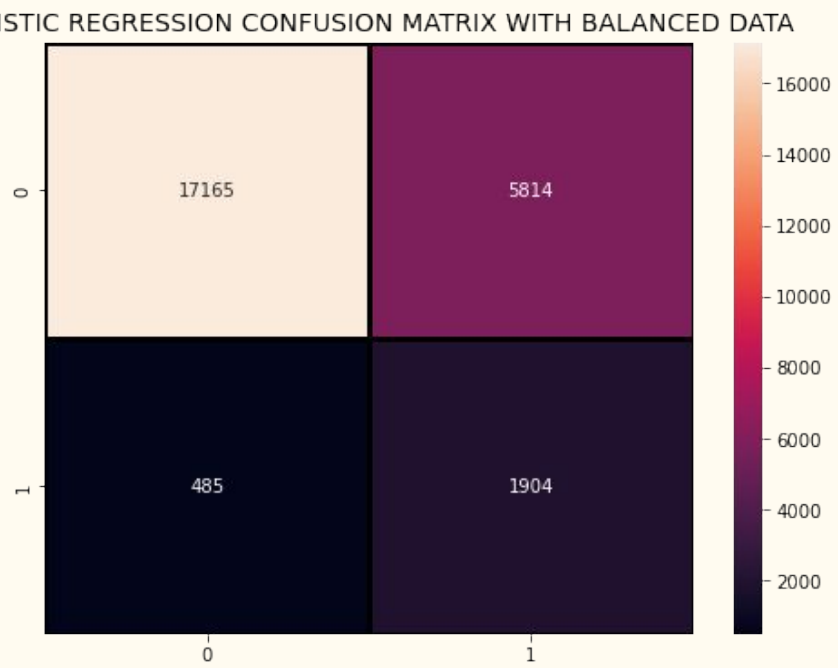
	Diabetes_target_mean	HighBP_target_mean	HighBP|Diabetes
253627	0.424895	0.652690	0.277325
195008	0.741493	0.654376	0.485215
10735	0.740414	0.652276	0.482954
111058	0.425296	0.292105	0.124231
159442	0.424939	0.652619	0.277324

Classification models

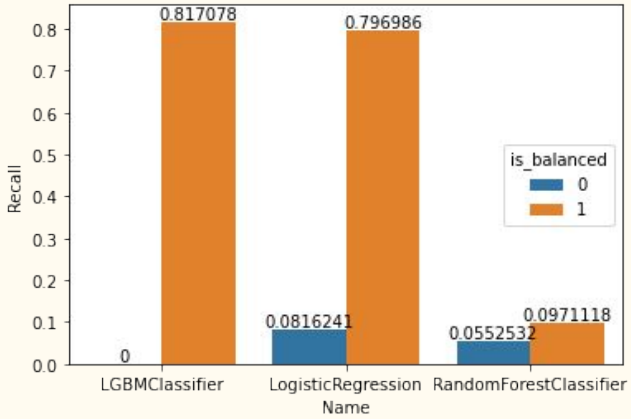
- We used LightGBM, Logistic Regression and Random Forest for this classification problem.
- For every Machine Learning algorithm, we trained on both balanced data set and imbalanced data set.
- We used GridSearchCV to search the best parameters for each model.
- Sensitivity was the most important factor while evaluating the models, hence we chose accuracy and recall as our evaluation metrics.

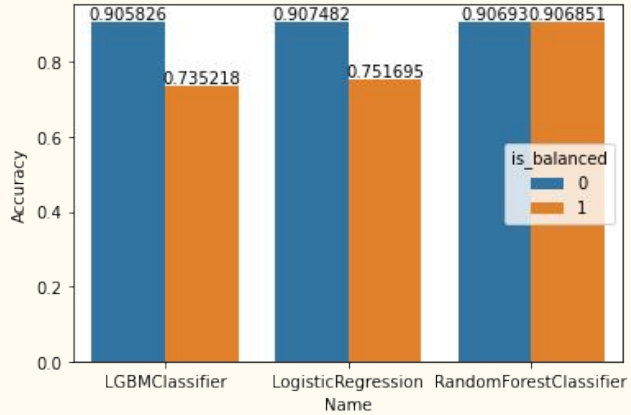
Confusion Matrix with Logistic Regression





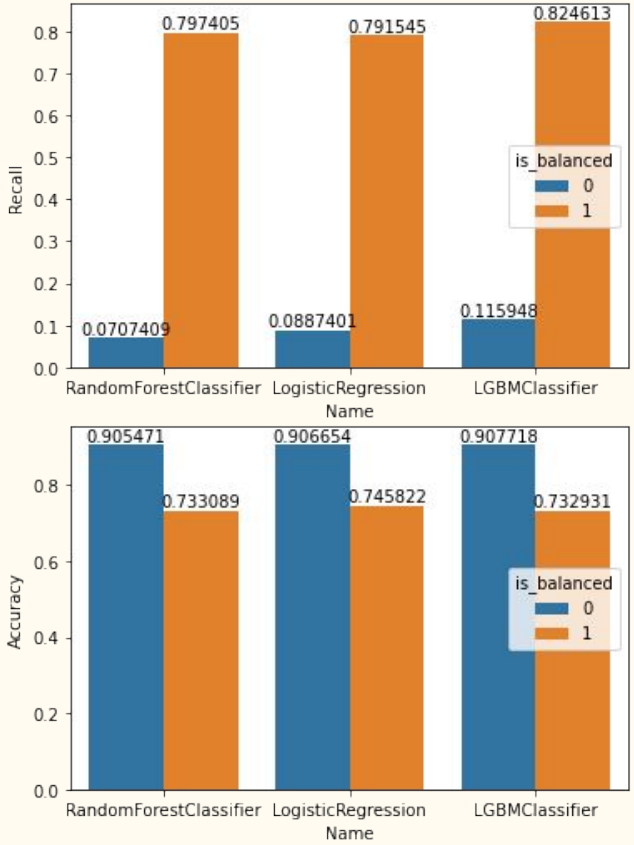
Final Evaluation : Accuracy and Recall





- Since we had fewer samples in the positive class, we under-sampled the over-represented label. Accuracy on balanced data set had dropped significantly. However, as a trade off, we could have a higher recall score.
- Even after using balanced data to train all the models, the recall score of Random Forest Classifier did not improve much.

Final Evaluation : Accuracy and Recall



- With a more complicated feature engineering strategy, we did not improve most of the models' performance except Random Forest.

Conclusion

- Since, the dataset is related to medical records, our main goal was to prevent avoiding any true positives.
- Sensitivity / probability of detection was utmost priority in this case, hence, accuracy and recall were chosen as evaluation metric.
- Although balancing data improved generalization of Logistic Regression and LightGBM models, it didn’t improve the performance of Random Forest Classifier.
- Feature engineering, on the other hand, improved the performance of all models, with LightGBM being the highest and Random Forest Classifier being the lowest.
- We started our project with the business problem of how to do early detection of heart disease in order to save nearly USD $15 billion in GDP, and in the end, we can say that our model can actually be used at a mass scale in order to detect heart disease at an early stage and thus in return help the researchers to stop/prevent the future cases of heart diseases.

Recommendations

- Medical experts can actually concentrate on using our proposed machine learning model to improve the heart disease-based clinical data analysis.
- Certain features (like HealthCare and DoctorCost), which aren't contributing to our 'target' column, could be eliminated while collecting data to save time in the future.

Thank You!!!