2024

UNIVERSITY OF LONDON

ST3189 Machine Learning Coursework Report

UOL STUDENT ID:

210500789

Table of Contents

Heart Disease Health Indicators Dataset[1]

Dataset variables[2]

	Name	Description of Variable
1	HeartDiseaseorAttack	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
2	HighBP	Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional
3	HighChol	Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high?
4	CholCheck	Cholesterol check within past five years
5	BMI	Body Mass Index (BMI) Continuous variable
6	Smoker	Have you smoked at least 100 cigarettes in your entire life?
7	Stroke	(Ever told) you had a stroke.
8	Diabetes	(Ever told) you have diabetes. 0 for no, 1 for pre-diabetes, 2 is for yes
9	PhysActivity	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
10	Fruits	Consume Fruit 1 or more times per day
11	Veggies	Consume Vegetables 1 or more times per day
12	HvyAlcoholConsump	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)
13	AnyHealthcare	Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service?
14	NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?
15	GenHlth	A Likert scale for the following question. Would you say that in general your health is: From 1 (Excellent) to 5 (Poor)
16	MentHlth	For how many days in the past 30 days was your mental health poor? (Includes stress, depression and emotional problems)
17	PhysHlth	For how many days in the past 30 days was your physical health poor? (Includes physical illness or injury)
18	DiffWalk	Do you have serious difficulty walking or climbing stairs?
19	Sex	Indicate sex of respondent
20	Age	Thirteen-level age category
21	Education	What is the highest grade or year of school you completed?
22	Income	Annual Household Income

[1] Dataset can be found on Kaggle. https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset

[2] Details on the dataset variables can be found on CDC website. https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

The dataset was derived from the Behavioral Risk Factor Surveillance System which is a collection of responses done through a telephone survey used to collect information from United States (U.S.) residents with regards to the respondent's demographic, behaviors that affect their health and any chronic health conditions.

1. Unsupervised Learning - Research Question

Can we identify any clusters or subgroups of patients based on their health indicators and habits with the use of unsupervised learning techniques?

Analysis

Principal Component Analysis (PCA)

Having a quick look at the data, with many categorical variables and numerical variables with various means and variances, we scaled the numerical variables through standardization. This ensures that each feature contributes equally to PCA, as it would prevent features like BMI, from dominating the variance, affecting the PCA calculations.

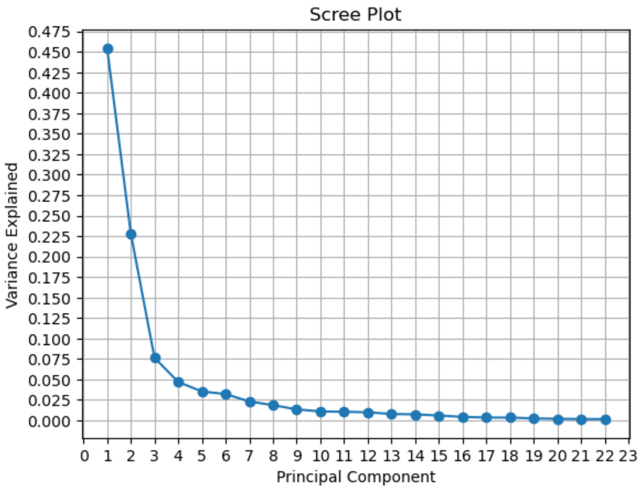
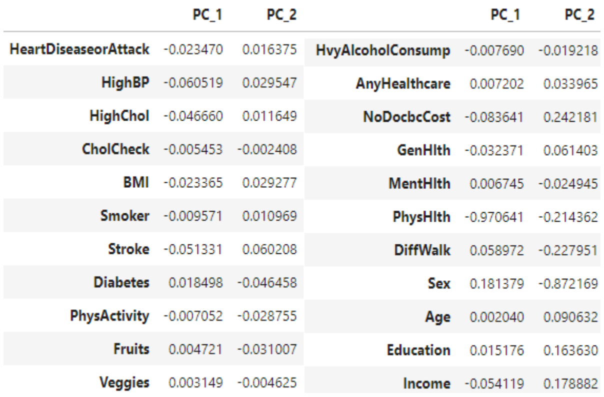


Figure 1. Principal Component Scree Plot



	PC_1	PC_2		PC_1	PC_2
HeartDiseaseorAttack	-0.023470	0.016375	**HvyAlcoholConsump**	-0.007690	-0.019218
HighBP	-0.060519	0.029547	**AnyHealthcare**	0.007202	0.033965
HighChol	-0.046660	0.011649	**NoDocbcCost**	-0.083641	0.242181
CholCheck	-0.005453	-0.002408	**GenHlth**	-0.032371	0.061403
BMI	-0.023365	0.029277	**MentHlth**	0.006745	-0.024945
Smoker	-0.009571	0.010969	**PhysHlth**	-0.970641	-0.214362
Stroke	-0.051331	0.060208	**DiffWalk**	0.058972	-0.227951
Diabetes	0.018498	-0.046458	**Sex**	0.181379	-0.872169
PhysActivity	-0.007052	-0.028755	**Age**	0.002040	0.090632
Fruits	0.004721	-0.031007	**Education**	0.015176	0.163630
Veggies	0.003149	-0.004625	**Income**	-0.054119	0.178882

Figure 2. Principal Component Loadings table

From Figure 1, it can be seen that a significant portion of the variance is captured in the first 2 Principal Components. The 1st Principal Component (PC) has an explained variance of 45.5% and the 2nd PC has an explained variance of 22.8%. 68.3% of the variance is captured just from the first 2 Principal Components.

From Figure 2, it displays the loadings for PC 1 and PC 2 of features that have an absolute value of loading of more than 0.1. From PC 1, PhysHlth is the dominant feature with a loading of -0.970641. This means the variance captured in PC 1 is mostly explained by the number of days from the past 30 days that the respondent was in poor physical health. From PC 2, Sex is the dominant feature with a loading of -0.872169. This means the variance captured in PC2 is mostly explained by the Gender of the respondent.

K-Means Clustering

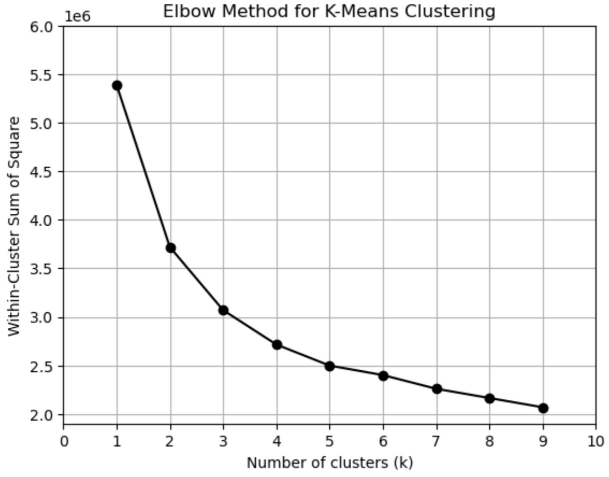


Figure 3. Elbow Method Plot

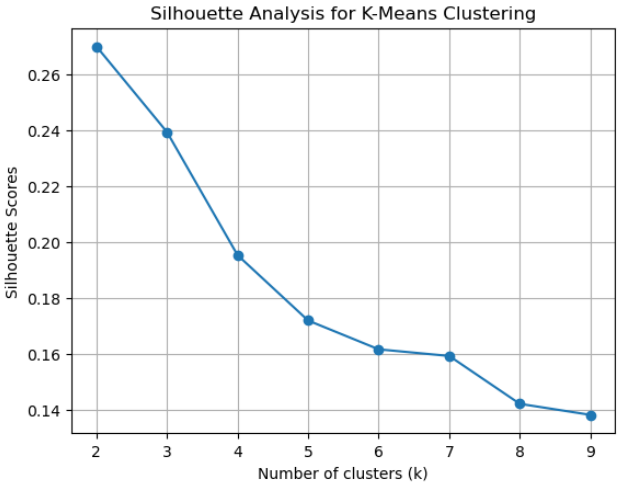


Figure 4. Silhouette Method Plot

Before we proceed with the K-means Clustering model, we will first need to determine the optimal number of clusters to use for our model. To determine the optimal number of clusters, we will be using the Elbow Method. Figure 3 displays a plot of the Within-Cluster Sum of Square at each total number of clusters, which is needed for the Elbow Method. However, from the plot there is no point on the graph where there exists a steep decline followed by a gentle decline, which could signify a clear elbow. Since we are unable to identify a clear elbow from the graph, we will need to proceed with Silhouette Analysis to determine the optimal number of clusters (K). For Silhouette Analysis, we calculated the silhouette score at each value of K. We repeated the calculation through 5 different random states and took the average silhouette score from those 5 random states to receive a more accurate estimation of the silhouette score at each value of K. From Figure 4, it is clearly shown that the optimal number of clusters is 2. Thus, we will be using K = 2 for the optimal number of clusters.

Hierarchical Clustering

Hierarchical Clustering can be a very computationally expensive clustering method and a few ways to decrease computational cost could be through dimensionality reduction through PCA or through sampling. However, the dataset contains many categorical features which would affect the results for PCA. Thus, we will be using a sample of 10% of the dataset.

Findings

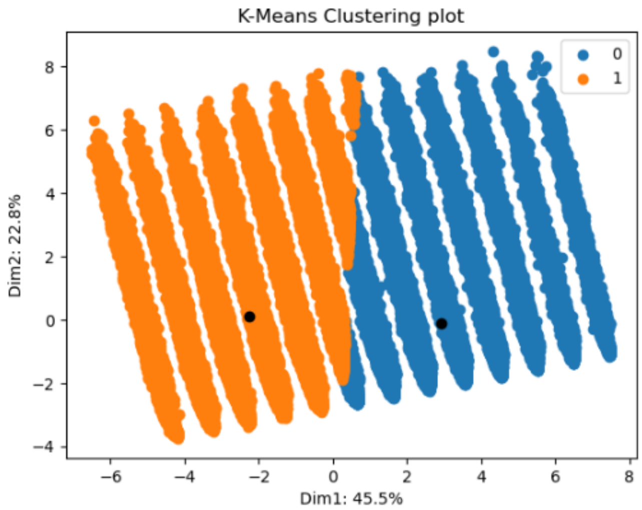


Figure 5. K-Means Clustering Plot

Figure 5 displays the cluster plot derived from my K-means Clustering model. It is apparent that the points in the plot segregate into 13 separate lines. This is most likely due to the age feature which has 13 levels. The presence of this categorical variable makes it difficult for the K-means clustering model to accurately calculate distance metrics. The presence of many categorical features in the data make it difficult to cluster accurately with K-means Clustering.

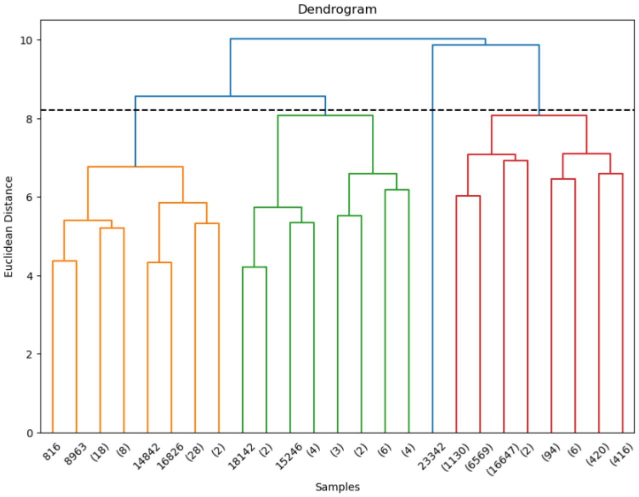


Figure 6. Hierarchical Clustering Dendrogram

Figure 6 displays the hierarchical clustering dendrogram where the linkage method used is Average. The point where the dotted line cuts across was chosen to be the threshold where the data in the dendrogram is separated into clusters as the points where the dotted line cuts across appear to be where the Euclidean distance is the longest. Through hierarchical clustering, the data is separated into 4 clusters.

For the dataset used, the best choice of clustering would be hierarchical clustering due to the number of categorical features in the dataset.

2. Classification – Research Question

Heart disease is the **leading cause of death** for men, women, and people of most racial and ethnic groups in the United States (Centers for Disease Control and Prevention, 2023).
According to Centers for Disease Control and Prevention (CDC), high blood pressure, high cholesterol and smoking put individuals at a higher risk of heart disease. Various medical conditions and certain lifestyle habits can also place individuals at a higher risk of heart disease.[3]

Research Question:
With the use of classification machine learning models, can we predict whether an individual has heart disease based on their health indicators and habits.

Analysis

For our analysis, we will be using Logistic Regression, K Nearest Neighbors, Decision Tree Classifier and Random Forest Classifier. To compare which model is the best fit for our prediction, we will be comparing the model's accuracy, precision, F1-score and ROC Curves. The dataset will be split into train set and test set with test size of 20% of the data. To keep the comparison fair, we will be setting 'random_state' for all models, if possible, to '2024'.

Logistic Regression

The accuracy of the model using default parameters is 0.9092. We proceeded with hyperparameter tuning to attempt to increase the accuracy of the model. From our hyperparameter tuning, the best parameters are the 'C' regularization parameter set to 0.1 and solver parameter set to sag. The max number of iterations was also increased to 500 from the default value of 100 due to the default model reaching the limit of 100 iterations with the default solver. Using the best parameters, the accuracy score of the model increased to 0.9093.

[3] Information taken from CDC. https://www.cdc.gov/nchs/nvss/leading-causes-of-death.htm

K-Nearest Neighbors

Before building the model, we first had to identify the optimal value of K for the K-Nearest Neighbors model. To figure out the optimal number of K, the model was put in a for loop where the model would run through values of K from 1 to 15 and we would then take the model with the highest accuracy. Figure 7 is a line graph of all the test scores at each value of K. The optimal value of K is 12 neighbors.

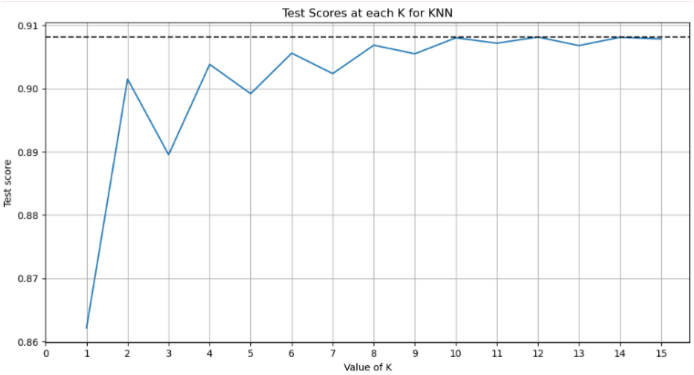


Figure 7. Test Scores at each K

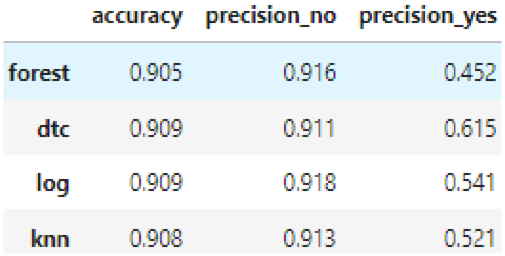
Decision Tree Classifier

The accuracy of the model using default parameters is 0.8523. We proceeded with hyperparameter tuning to attempt to increase the accuracy of the model. From our hyperparameter tuning, the best parameters are the max depth of the tree is set to 4, the minimum number of leaves remains at the default value of 1 and the minimum number for the leaf to be allowed to split remains at the default value of 2. Using these best parameters, the accuracy score of the model increased to 0.9091.

Random Forest Classifier

The model was built using the default parameters. Hyperparameter tuning was not implemented as the dataset is large and would be too computationally expensive.

Findings



	accuracy	precision_no	precision_yes
forest	0.905	0.916	0.452
dtc	0.909	0.911	0.615
log	0.909	0.918	0.541
knn	0.908	0.913	0.521

Figure 8. Model Scores

Figure 8 displays the accuracy score for all the models used. From the table, the Decision Tree Classifier and Logistic Regression models have the highest accuracy scores of 0.909. Figure 8 displays the precision scores for all the models used. From the table, the Logistic Regression model is the best for predicting people who have no heart disease and Decision Tree Classifier is best for predicting people who have heart disease.

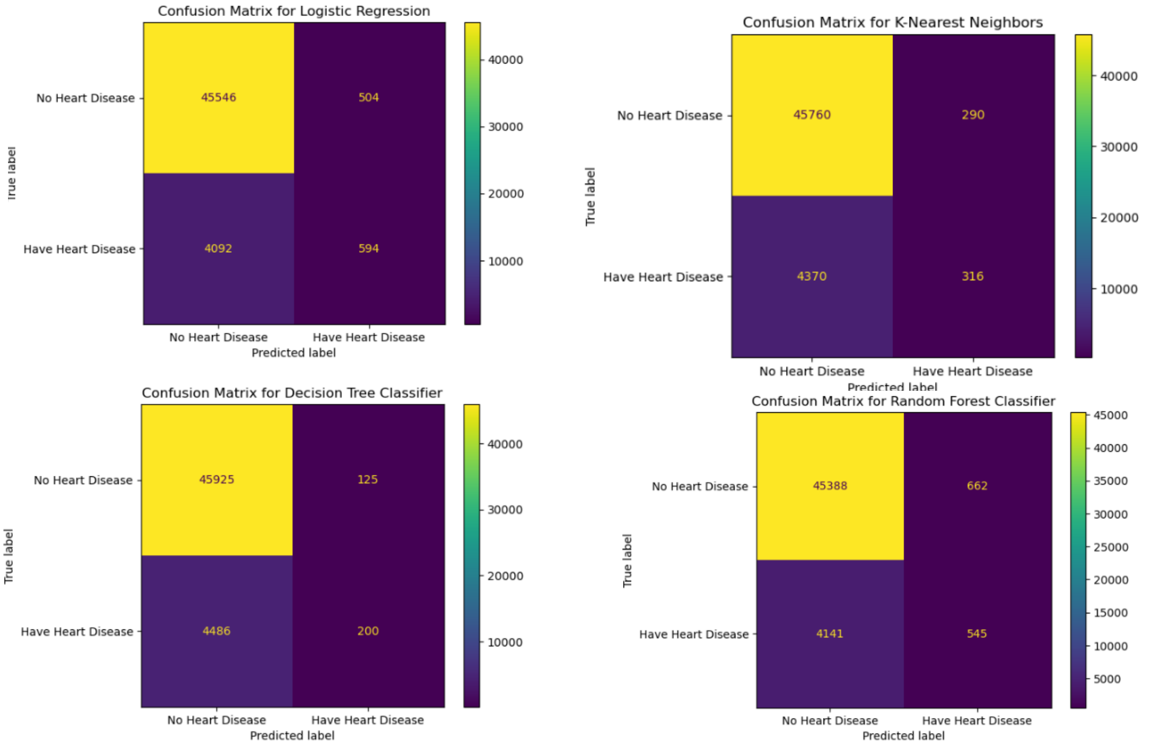


Figure 9. Confusion Matrix of all Models

However, in the case of heart disease, it will be better if we can predict a higher number of people who have heart disease instead of focusing on the precision of the prediction. Thus, we will be looking at the confusion matrix for all the models to see which model correctly predicts the greatest number of people who have heart disease. Figure 9 shows the Confusion Matrix for all the models used. Comparing the Confusion Matrixes of all the models, the Logistic Regression model is best for correctly predicting the greatest number of individuals who have heart disease, correctly predicting 594 cases.

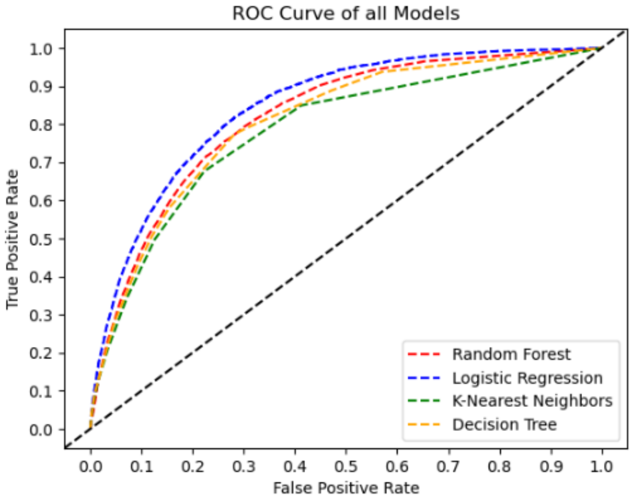


Figure 10. Roc Curve of all Models

Figure 10 displays the ROC curve for all the models used. From the ROC curves, it can be clearly seen that Logistic Regression has the biggest area under the curve compared to the ROC curve of the other models. This means that the Logistic Regression model performs better for the dataset used compared to the other models.

Conclusion:

From our findings, we can deduce that the Logistic Regression model is the best fit for the dataset used. Logistic regression has the highest area under the curve from the ROC curves which indicates that it performs better compared to the other models. From the confusion matrixes, the logistic regression model is able to correctly predict the highest number of individuals who have heart disease.

Housing Price prediction dataset[4]

Dataset variables

	Name	Description of Variable
1	price	Price of the house
2	area	Area of the house (in sq. feet)
3	bedrooms	Number of bedrooms
4	bathrooms	Number of bathrooms
5	stories	Number of floors
6	mainroad	Is there a main road connected to the house?
7	guestroom	Does the house have a guest room?
8	basement	Does the house have a basement?
9	hotwaterheating	Does the house have a hot water heating system?
10	airconditioning	Does the house have an air conditioning system?
11	parking	Number of parking spaces in the house
12	prefarea	Is the house located in a preferred area?
13	furnishingstatus	Furnishing status of the house (Furnished, Semi-furnished or Unfurnished)

3. Regression – Research Objective

Our objective is to predict the price of houses based on the variables shown above and by how much these variables affect the housing price.

With housing becoming more affordable in the United States in recent years, we want to look at the various factors and variables that will affect the housing prices.

The Y variable will be the price and the X variables will be the rest. Since 'area' is the only linear variable from the X variables, we don't need to do any scaling on the X variables, however we would still need to conduct encoding on the categorical variables. For this case, we conducted label encoding. Categorical variables with yes/no encoded to 1/0 respectively and for the 'furnishingstatus' variable, furnished encoded to 0, semi-furnished encoded to 1 and unfurnished encoded to 2. We will be setting train size as 80% of the data and test size as 20% of the data.

[4] Dataset can be found on Kaggle. https://www.kaggle.com/datasets/harishkumardatalab/housing-price-prediction/data

Analysis

Linear Regression

Linear regression is the first model we will be using to predict the housing price. ‘price’ will be the dependent variable and the rest of variables will be the independent variables. There are no hyperparameters needed to tune for the linear regression model, thus the default model function is used.

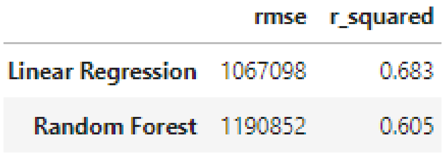
Random Forest Regression

The second model we will be using is the Random Forest regression model. Using the default parameters from the model function, the r-squared of the model is 61.0%. To try to improve the performance of the model, we will be doing hyperparameter tuning. From our hyperparameter tuning testing, we have discovered that the best parameters are max depth of the trees set to 10, the minimum number of leaves set to the default value of 1, the minimum number for the leaf to be allowed to split set to the default value of 2 and the number of decision trees to be built set to 50. Using these parameters, the r-squared of the model drops to 60.5%. This could suggest that there is overfitting in the original forest regressor model. When assessing the r-squared of the default model on the train set, the Mean Squared Error (MSE) of the train set is significantly lower than the MSE of the test set which suggests the presence of overfitting. Also, the r-squared of the train set is 93.4% while the r-squared of the test set is 61.0%. Thus, we will be using the best parameters for our Random Forest regressor model.

XGBoost Regressor

The third model we will be using is the XGBoost Regressor model. Using the default parameters from the model function, the r-squared of the model is 0.540. We will be doing hyperparameter tuning to try to improve the performance of the model. From our hyperparameter tuning testing, the best parameters are the learning rate of the model set to 0.1, max depth of the tree set to 3 and number of trees to build set to 100. Using these parameters, the r-squared of the model improves significantly to 0.648. When comparing the Root Mean Squared Error (RMSE) of the model before and after, the RMSE value improves from 1,284,902 to 1,124,488. Thus, we will be using the best parameters for our XGBoost regressor model.

Findings



	rmse	r_squared
Linear Regression	1067098	0.683
Random Forest	1190852	0.605
XGBoost	1124488	0.648

Figure 11. RMSE and R-Squared for all models

Figure 11 is a table which displays the RMSE and r-squared scores for all the models. From the table, it is apparent that the linear regression model has the best model fit for our dataset as it has the lowest RMSE value of 1,067,098 as well as the highest r-squared of 0.683.

References

Centers for Disease Control and Prevention (CDC). (2023, May 15). *Heart Disease Facts.* https://www.cdc.gov/heartdisease/facts.htm

Centers for Disease Control and Prevention (CDC). (2016, August 23). *Behavioral Risk Factor Surveillance System 2015 Codebook Report.* https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

Centers for Disease Control and Prevention (CDC). (2023, March 21). *Know Your Risk for Heart Disease*. https://www.cdc.gov/heartdisease/risk_factors.htm

World Health Organization (WHO). (2021, June 11). *Cardiovascular diseases (CVDs)* https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Statistica Research Department. (2023, September 28). *Housing affordability index (fixed) in the United States from 2000 to 2022.* https://www.statista.com/statistics/201568/change-in-the-composite-us-housing-affordability-index-since-1975/

National Association of Realtors. *Housing Affordability Index.* https://www.nar.realtor/research-and-statistics/housing-statistics/housing-affordability-index#:~:text=The%20Housing%20Affordability%20Index%20measures,recent%20price%20and%20income%20data.

YCharts. *US Fixed Housing Affordability Index (I:USFHAI)* https://ycharts.com/indicators/us_fixed_affordability_index

Conor Dougherty. (2024, March 27). *America's Affordable Housing Crisis* https://www.nytimes.com/2024/03/27/briefing/affordable-housing-crisis.html

Alex Teboul. *Heart Disease Health Indicators Dataset.* https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset

HARISH KUMARdatalab. *Housing Price Prediction.* https://www.kaggle.com/datasets/harishkumardatalab/housing-price-prediction/data