UNIVERSITY OF LONDON

LSE

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

AUTHOR: BRENDON TREVIN THAPARARATNAM

PAGE COUNT: 10 PAGES
(EXCLUDING COVER PAGE, TABLE OF CONTENTS AND APPENDIX)

Table of Contents

BACKGROUND / CAVEAT

Please refer the appendix for glossary of terms and abbreviations

Links for the datasets used in each task have also been included in the appendix

A train-test split of 75-25% of the data was carried out for all models where necessary

Both grid search and randomized search were conducted for hyperparameter tuning wherever possible. The results mentioned in the report are the best out of the two.

TASK 1: UNSUPERVISED LEARNING

INTRODUCTION

Unsupervised learning is where models are trained using unlabelled data in order to find patterns within the dataset. This can be done for the purpose of ***clustering*** where we find similarities and form groups, ***association*** where we find the correlation between data and ***dimensionality reduction*** where the number of features is reduced while preserving data integrity.

The wholesale customer dataset (Abreu, N., 2011) was used for this task. It consists of 8 variables which collectively describes the annual spending (in monetary units) of clients of a wholesale distributor on a diverse range of products.

RESEARCH QUESTIONS (RQ)

1. How does Milk distribution differ between Hotels/Restaurants/Café (HoReCa) and retailers?
2. Is there a correlation between Grocery, Milk and Detergents/Paper products?
3. How does the annual mean spending differ between Lisbon, Oporto and Other Regions?

LITERATURE REVIEW

A study by (Popovic, 2009), showed that milk products are mainly distributed through retailors and only a very small amount through food service (HoReCa). Retail stores have a distribution participation share of over 99% for fresh milk, 97.2% for yoghurt and 93.8% for UHT milk.

(Abdullah et al., 2016) states that based on a study done in Portugal in 2011, there is a strong positive correlation between the annual spending on Grocery and Detergents/Paper products and also between Grocery and Milk products.

The study further states that there is no significant difference in the annual mean spending on Fresh, Milk, Grocery, Frozen, Detergents and Paper products between Lisbon, Oporto and Other regions.

EXPLORATORY DATA ANALYSIS

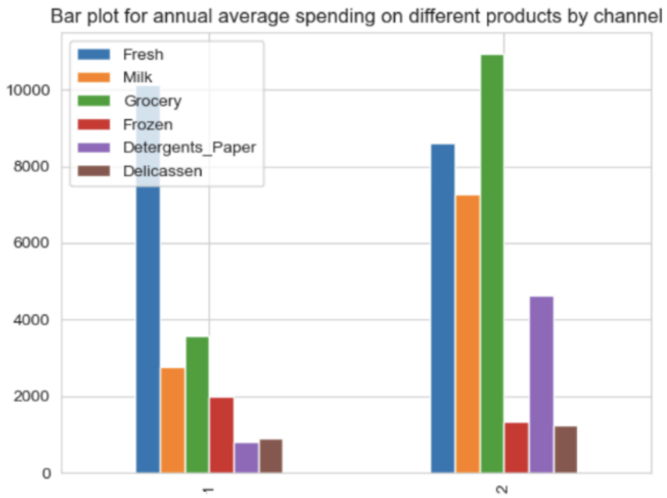


Figure 1

RQ1: Figure-1 shows that Channel 1 (HoReCa) has significantly less average spending on Milk than compared to Channel 2 (Retail). Suggesting a trend of retailers purchasing significantly more Milk than HoReCa distributors from wholesalers.

RQ2: Figure-2a shows that Grocery and Detergents_Paper have a positive linear relationship with a correlation of 0.85(figure-3, appendix). Similarly, Grocery and Milk(Figure-2b) have a positive linear relationship with a correlation of 0.74. Hence these 3 products indeed have a strong positive correlation. Apart from that, Channel 2 has more exorbitant and volatile prices than compared to Channel 1. However, from the density of points Channel 1 can be seen to have higher sales.

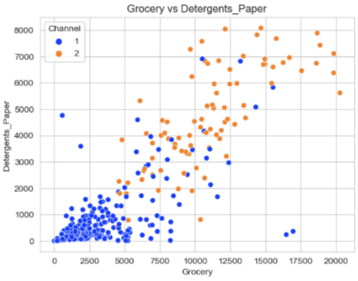


Figure 2a

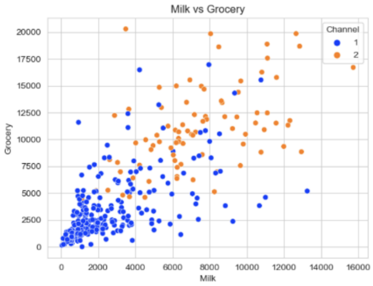


Figure 2b

RQ3: According to Figure-4, the mean annual spending across all regions follow a similar trend for all product ranges. Fresh, Grocery and Milk show the highest averages throughout. Hence there is no regional discrepancy between Lisbon, Oporto and Other.

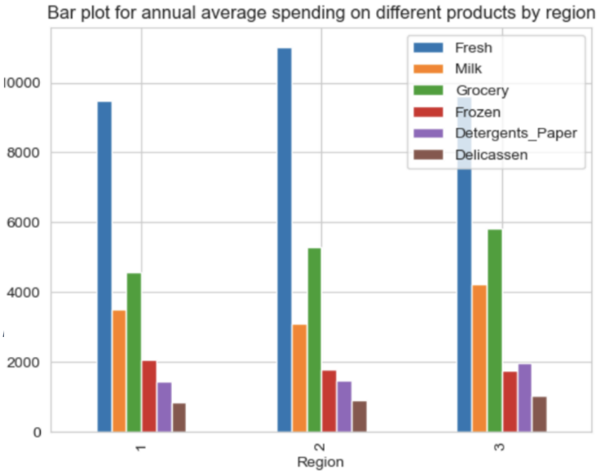


Figure 4

PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a method of dimensionality reduction that transforms a larger set of variables into smaller ones known as Principal Components (PCs) which increases interpretability while retaining as much information as possible. PCA is done to reduce complexity and size for easier visualizations and more efficient running times of machine learning algorithms.

Methodology

Since PCA is quite sensitive to larger variances of initial variables we standardize the data by scaling them so that biased results are avoided.

PCs are built such that the first PC contains the largest variance and as we ascend every sequential PC contains even less so. PCA is done using the concept of eigen decomposition, where eigen vectors show the direction of spread and eigen values give the relative importance of these directions. As such the eigen vectors give the PCs and eigen values depict the explained variance retained by these PCs. Greater the explained variance of a PC the higher the information retained. By sorting the eigen vectors using their eigen values/explained variance scores a ranking can be given to the PCs which allows us to select the optimal number of PCs.

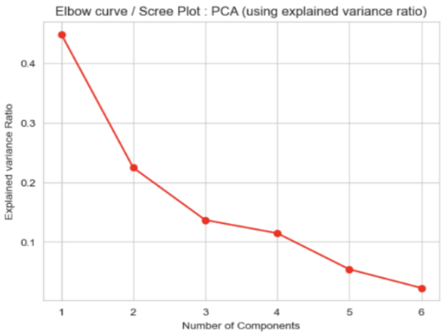
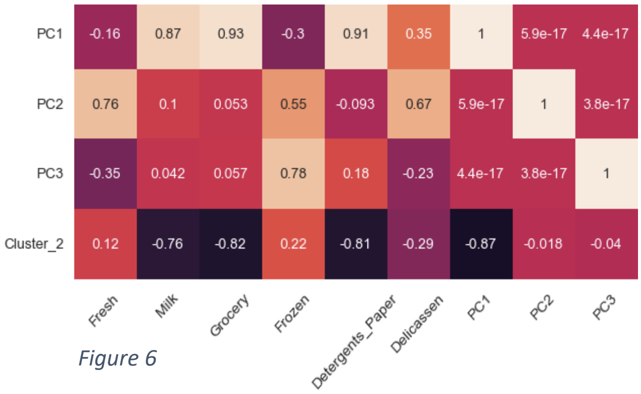


Figure 5a

The scree plot (Figure-5a) depicts the proportion of explained variance/ eigen values of PCs in descending order. Using the elbow method, we can find the optimum number of PCs. The elbow can be seen to occur at the 3rd PC. The first PC explains ~45% of the variance, the second ~22% and the third ~13%. Hence the first 3 PCs together explain roughly 80% of the variation in the data. This alone is sufficient to conclude that 3 is the optimal number of PCs

However, diving further, Figure-5b(appendix) shows Kaiser's rule where we select the number of PCs above the eigen value of one as long as we have a cumulative explained variance of at least 80% (Figure-5c, appendix). Hence as we achieve a cumulative explained variance of 80% only at the 3rd PC and not at the 2nd, we conclude that 3 is our optimal number of PCs.



	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	PC1	PC2	PC3
PC1	-0.16	0.87	0.93	-0.3	0.91	0.35	1	5.9e-17	4.4e-17
PC2	0.76	0.1	0.053	0.55	-0.093	0.67	5.9e-17	1	3.8e-17
PC3	-0.35	0.042	0.057	0.78	0.18	-0.23	4.4e-17	3.8e-17	1
Cluster_2	0.12	-0.76	-0.82	0.22	-0.81	-0.29	-0.87	-0.018	-0.04

Figure 6

We now analyze the loadings of each feature on each PC. This gives us an idea of the correlation between the features and the PCs. As per the correlation matrix (Figure-6), it can be seen that our 3 PCs are uncorrelated with eachother. Milk, Grocery and Detergents_Paper have a very high correlation with

PC1 which is expected as it contains most of the variance. While Fresh, Frozen and Delicassen have a distinct correlation with PC2. Whereas PC3 only shows a distinct correlation with Frozen. Thus inter-component correlation is low while intra-component correlation is high.

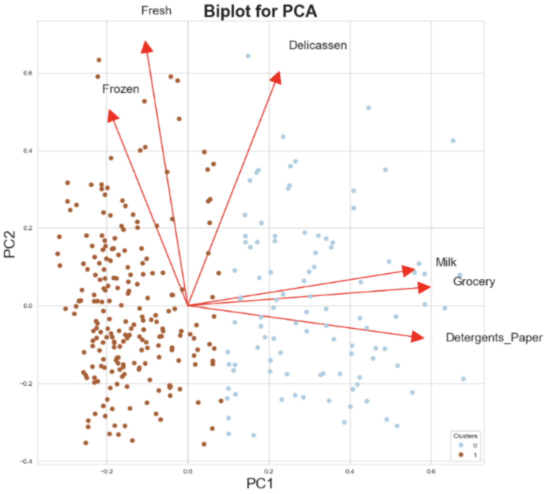


Figure 7

Interpretation of PCs are easier when visualized in a 2D manner. Hence we analyze a biplot for the first 2 PCs as they contain most of the variation in the data and have significant correlation with all features.

As per Figure-7, we see that Milk, Grocery and Detergents_Paper contribute more to PC1 than PC2 and has a strong positive correlation with it. Fresh, Delicassen and Frozen contribute more to PC2 than PC1 and also has a strong positive correlation with it. As the lengths of the arrows are all large, all features show large contributions to their respective PCs

Frozen and Fresh have a small negative correlation with PC1. While Detergents_Paper has a very small negative correlation with PC2.

Moreover, it is visible that Milk, Grocery and Detergents_Paper are uncorrelated with Fresh and Frozen due to them being at obtuse angles.

Milk, Grocery and Detergents_Paper have a high correlation between eachother as they are at acute angles. Similarly, Frozen, Fresh and Delicassen have a high correlation among eachother.

K-MEANS CLUSTERING

K-Means clustering is defined as an iterative algorithm that groups unlabelled data into different clusters such that the sum of distances between data points within a cluster and their respective cluster centroids is minimized.

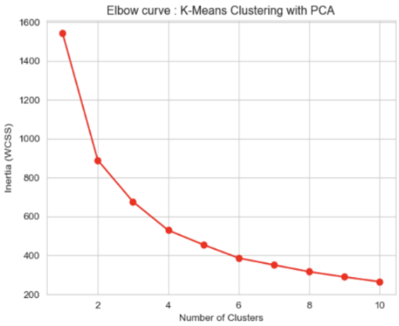


Figure 8

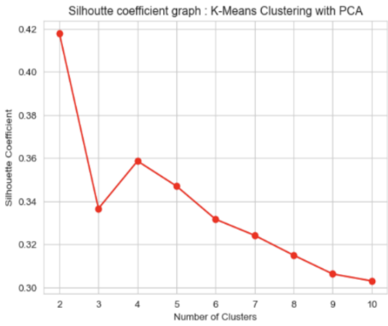


Figure 9

In order to identify the ideal number of clusters we make use of the elbow curve and silhoutte coefficients. Using Figure-8 we apply the elbow method to identify the point at which the within sum of squares (WSS or inertia) is minimized. The elbow bend occurs at two clusters.

In addition we can also use the silhoutte score curve (Figure-9) which evaluates the goodness clustering using inter-cluster and intra-cluster distance. The peak occurs at two clusters. Hence, the optimum number of clusters was determined to be 2 for the given dataset.

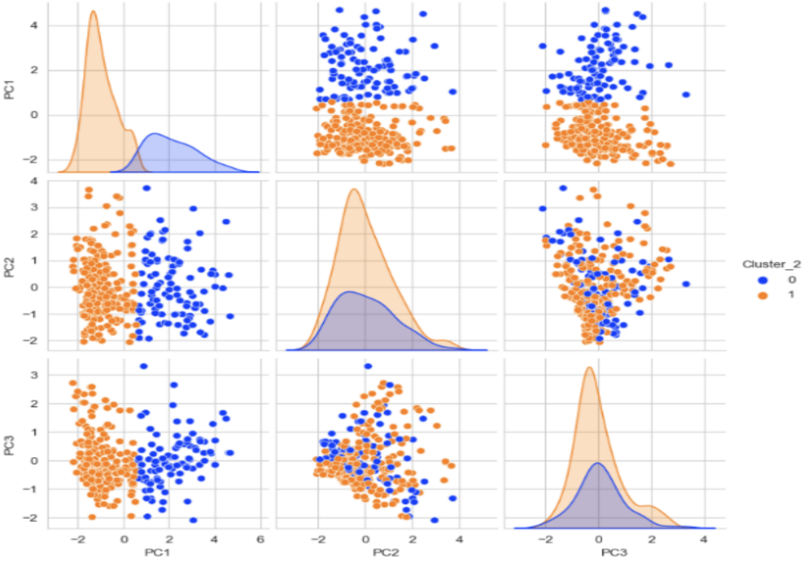


Figure 10

As shown in Figure-10, distinct clusters are visible between PC1-PC2 and also PC1-PC3. No clusters can be found between PC2-PC3.

TASK 2: REGRESSION

INTRODUCTION

Supervised learning is the use of labelled datasets as inputs to train algorithms in order to classify or predict outcomes of target variables.

Regression analysis is a statistical method used to determine the strength and relationship between one continuous dependent variable (target variable) and one/many independent variables

The Boston Housing Price dataset (US Census Service, 1978) was used for this task. It consists of 14 variables which collectively describes the housing and living conditions of the Boston metropolitan area. The median value of owner-occupied homes (MEDV) was considered as the target variable. A regression model to predict the MEDV was expected to be built.

RESEARCH QUESTIONS (RQ)

1. Does number of rooms per dwelling have a significant impact on housing prices?
2. How does age of a house affect its market price?
3. Do air pollution levels affect the prices of housing?

LITERATURE REVIEW

According to (Selim, 2008) a higher number of rooms and house size lead to higher prices. However, the magnitude of change in prices according to room count fluctuations depended on many other factors such as locational characteristics (urban versus rural areas), quality of amenities (central heating, wall hung gas boilers etc.), floor types and other structural characteristics whose effects may vary between 14%-156%.

As per (Fan et al., 2006) the hedonic regression approach showed that house age plays a major role in determining the price of a house. The prices of houses between 5-10 years are less than that of 0-5 years by 8% and 12% respectively. Thus, the older the house the lower the price.

According to (Harrison & Rubinfeld, 1978) air quality plays an important role in house prices. As air pollution (NOX) levels increase the prices of houses generally fall. However, the elasticity of willingness to pay for low, middle and high income groups drops to 0.97, 0.94 and 0.90 correspondingly. Hence, as pollution reduces, the marginal valuation of air quality improvements recedes more quickly for lower income households than middle and high income households.

EXPLORATORY DATA ANALYSIS

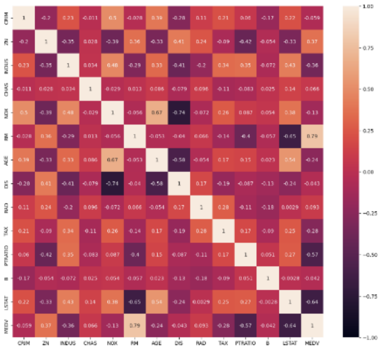


Figure 11

RM and LSTAT show a distinct correlation with MEDV (Figure-11). Furthermore, AGE, NOX, DIS, PTRATIO, INDUS, TAX and CRIM show a slight correlation with MEDV.

A very sparse number of houses were situated close to the Charles River. There were only 2 houses near the river where LSTAT was greater than 20%. Meaning, being situated closer to the river was considered a luxury in the Boston market.

The mean and median prices of houses closer to the river were higher in general. Especially the starting price of houses, which were significantly higher than houses far from the river (~$20,000 vs $12,000). Furthermore, there seems to be a lot of outliers for houses situated further away from the river. Hence this could lead to inaccurate predictions of prices of houses further away from the river. Prediction of prices would likely be more dependable for houses which are closer to the river.

RQ1: According to Figure-12 it is evident that there is a distinct upward trend between RM and MEDV. Hence as number of rooms increase the price of houses do tend to rise. Moreover, there seems to be some correlation between RM and other variables such as LSTAT and PTRATIO (Figure-11)

RQ2: As per Figure-13 there seems to be a negative relationship between AGE and MEDV. As AGE increases MEDV tends to fall quicker at first but towards the higher spectrum of the AGE-axis the rate of fall in price tends to reduce. Hence as age of the house increases the price of houses do tend to fall at varying rates.

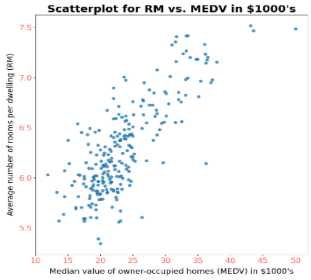


Figure 12

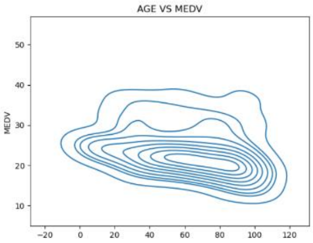


Figure 13

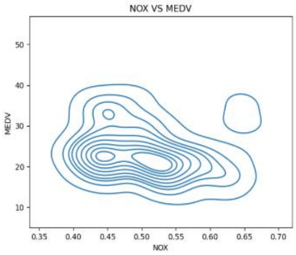


Figure 14

RQ3: Figure-14 shows a negative relationship between NOX and MEDV. As NOX levels rise the MEDV reduces. However, the slope is not drastic enough for it to cause a significant impact on prices. Hence, there is some relationship between air pollution levels and prices of houses.

MODELS

Multiple Linear Regression

"Multiple linear regression" is the modelling of many explanatory variables and a continuous response variable by fitting a linear equation to observed data. Ordinary Least Squares (OLS) method is used here in fitting the model. That is the model coefficients are estimated such that we minimize the residual sum of squares between observed and predicted targets by linear approximation. Given n observations, the model is given by:

$$Yi = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \ \varepsilon_i \ ; \ for \ i = 1,2, \ldots, n$$

Assumptions

Evan's (1996) significance value of threshold of Pearson correlation was used to assess feature selection and avoid multicollinearity: -

- 0.00-0.19: very weak
- 0.20-0.39: weak
- 0.40-0.59: moderate
- 0.60-0.79: strong
- 0.80-1.00: very strong

We assume that the data satisfies linear regression assumptions and the Gauss-Markov theorem: -

- Randomness- Data has been randomly sampled from the population
- Linearity- Linear relationship between response variable and explanatory variables
- No multicollinearity- The regressors are not perfectly correlated with each other
- Homoscedasticity- Error terms have constant variance
- Exogeneity- Regressors are not correlated with the error terms
- Error terms are independent themselves
- Error terms are normally distributed.

Methodology

Feature selection was carried out using the K-score which quantifies which features in our data contribute most to the target variable. Consequently, 10 out the 14 variables were selected as representative features ensuring there was no multicollinearity. Thereafter the model was built.

Analysis

- *R-squared (R^2)*: A statistical measure of fit that quantifies how much of the variation in the model is explained by the independent variables.
- *Explained variance score*: Similar to R-squared but it uses the biased variance to determine the fraction of variance explained. (In R-squared the raw sum of squares is used)
- *Mean Squared Error (MSE):* The average of the square of the difference between actual and predicted values
- *Root Mean Squared Error (RMSE):* The square root of MSE or the standard deviation of the residuals (prediction errors)

- *Mean Absolute Error (MAE):* The mean of the absolute difference between actual and predicted value

The explanatory variables explain 79.6% of the variance in the model. Hence, the model is satisfactory.

Decision Tree Regression

Decision trees are tree structure models which consist of a root node, decision nodes leaf nodes. The algorithm predicts the target variable using the tree representation where the leaf nodes represent classes and decision nodes represent attributes.

The algorithm used here is called ID3 by J. R. Quinlan which essentially does a greedy search in a top-down manner throughout the branches without backtracking. Decision trees are usually used for classification problems. However, we use it for regression aswell by replacing Information Gain with Standard Deviation Reduction.

Decision trees are not affected by multicollinearity hence we need not check for high correlation between independent variables.

Assumptions

- Initially the whole training dataset is considered the root
- Recursive distribution of records based on attributes
- If feature values are continuous, they are discretized

Analysis

The model initially showed an R^2 of 69.8% and an RMSE of 3.17.

We conduct hyperparameter tuning (Grid Search, Randomized Search, etc.) in order to find the most optimum parameters for the model and to also prevent overfitting.

Grid search proved to be the better tuning protocol. The main optimum parameters were: criterion=absolute_error and max_depth=6. Improvements were thus obtained with an R^2 of 71.8% and an RMSE of 3.07. Hence the model was considered satisfactory.

Random Forest Regression

Random forest regression is an ensemble learning method which uses multiple decision trees and averages the result to output a new result which often produce stronger predictions. Basically, it reduces the variance in trees by taking the average of test error estimates. Random forests are also immune to multicollinearity.

Analysis

The model initially showed an R^2 of 83.6% and an RMSE of 2.35.

Randomized search proved to be the better tuning protocol. The main optimum parameters were: criterion=absolute_error, bootstrap=True and max_depth=22. Post tuning results were an R^2 of 83.9% and an RMSE of 2.32. Thus, an extremely small improvement was made. Hence the initial model itself was considered satisfactory.

XG Boost Regression

Extreme Gradient Boosting is an ensemble learning algorithm used for regression and classification problems which attempts to predict a target variable by combining the estimates of weaker models. The weak learners (decision trees) are inserted one by one to the ensemble so that prediction errors from prior models are gradually corrected. This process is called boosting. A gradient descent algorithm is then used to minimize an arbitrary differentiable loss function through an iterative process as new models are added. The final result is a much stronger prediction.

All boosting algorithms are immune to multicollinearity.

Analysis

The model initially showed an R^2 of 86.0% and an RMSE of 2.16.

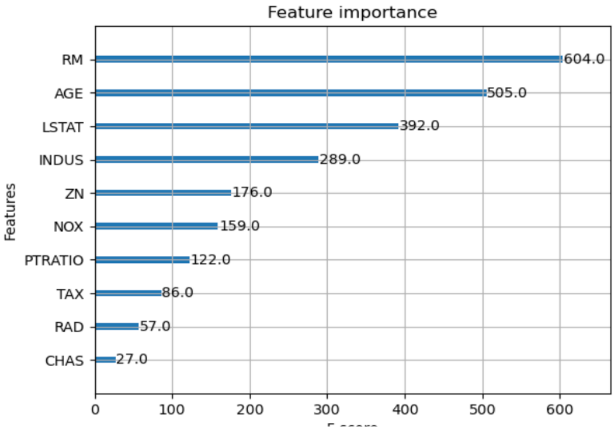


Figure 15

Randomized search proved to be the better tuning protocol. The main optimum parameters were: booster=gbtree, max_depth=11 and subsample=0.3. Post tuning results were an R^2 of 86.6% and an RMSE of 2.12. Thus, a marginal improvement was made. Hence the resulting model was considered highly satisfactory.

The feature importance graph (Figure-15) indicates how useful each feature was in the construction of the XGboost model by ranking all the features used using an F-score. It can be seen that the most useful features were RM, AGE, LSTAT, INDUS, ZN and NOX. This further solidifies the justifications to the above-mentioned research questions.

CONCLUSION

Performance metrics Models	Explained Variance Score		R-Squared (R^2)		Mean Squared Error (MSE)		Root Mean Squared Error (RMSE)		Mean Absolute Error (MAE)	
Multiple linear regression	0.805	-	0.796	-	6.79	-	2.601	-	1.95	-
Decision trees	0.700	0.725	0.698	0.718	10.07	9.41	3.17	3.07	2.43	2.22
Random forest	0.841	0.840	0.836	0.839	5.52	5.38	2.35	2.32	1.68	1.60
XG Boost	0.862	0.869	0.860	0.866	4.68	4.48	2.16	2.12	1.64	1.54

Figure 16

In Figure-14 the values on the left are prior to hyper parameter tuning and the values on the right are post hyperparameter tuning.

Explained variance score and R-Squared which is closer to 1 is an indication of a good model. Lower MSE, RMSE and MAE are indications of a good model.

The yellow highlighted values are the best scores of each performance metric. As such, it can be seen that the XG Boost regression model shows the best performance throughout; with 86.6% of the model's total variance being explained by the regressors and also having the lowest RMSE and MAE out of all the models with 2.12 and 1.54 respectively.

TASK 3: CLASSIFICATION

INTRODUCTION

Classification is a supervised learning algorithm where the model predicts the correct label of the categorical target variable for a given dataset.

The Heart Attack Prediction dataset was used for this task. It consists of 14 variables which collectively describe the cardiovascular well-being of each of the 303 patients. The vitals of each patient were recorded and then they were classified as having less chance of heart attack (0) or more chance of having heart attack (1). The aforementioned target variable was named output.

RESEARCH QUESTIONS (RQ)

1. Do women have more heart attacks at older ages than men?
2. Does heart rate have an effect on the probability of getting a heart attack?

3. Does the number of major vessels coloured by fluoroscopy have an association with the chances of getting a heart attack?

LITERATURE REVIEW

According to (Maas & Appelman, 2010), cardiovascular disease develops 7-10 years later in women than compared to men. This is mainly due to the hormonal changes caused by menopause. The exposure to endogenous oestrogens during the fertile period in premenopausal women delays the occurrence of atherosclerotic (narrowing of arteries) disease. However, after menopause, the levels of oestrogen reduce in turn increasing the risk of atherosclerosis. Therefore, as menopause occurs in the later years (around 45-55 years), this makes women more susceptible to heart disease later on in life.

(Perret-Guillaume et al., 2009) says that heart rate has a direct causality on cardiovascular morbidity and mortality. It was reported that a high maximum heart rate was associated with a higher risk of heart disease. An increase in heart rate by 10BPM had an increased risk of cardiac death by atleast 20%. Furthermore, this trend was stronger in males than in females

A study by (Mahmood & Kuppa, 2010) showed that the number of major vessels colored by fluoroscopy and chest pain type were critical biomarkers of heart disease. Cardiac catheterization checks for blockages in the four major arteries: aorta, superior vena cava, inferior vena cava and pulmonary artery. When there is no color it is an indicator of a blockage. The number of arteries blocked has a direct effect on the risk of having heart failure.

EXPLORATORY DATA ANALYSIS

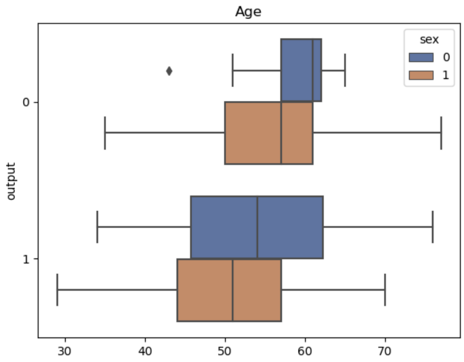


Figure 17

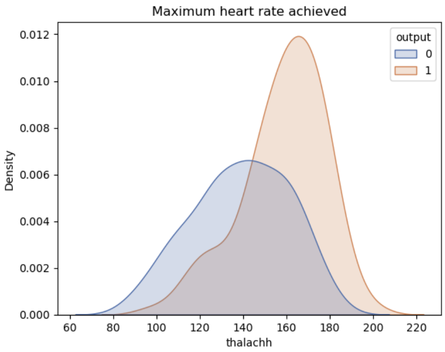


Figure 18a

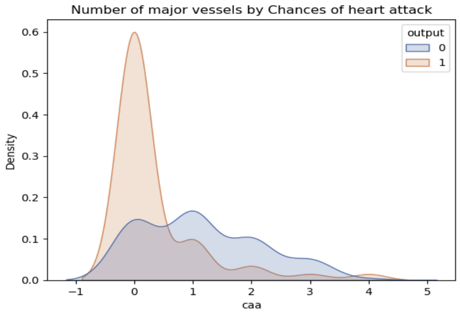


Figure 19

RQ1: According to Figure-17, it can be seen that out of those who have a high chance of heart attacks females (sex=0) have a larger Inter Quartile Range (IQR) for age than males (sex=1). Moreover, the median age for females is higher than males in both cases. Females who have a higher chance of heart attack having a higher median age is likely due to hormonal changes and menopause in women.

RQ2: As per Figure-18a, it can be seen that the maximum heart rate for patients with less chance of heart attack was 140BPM while for patients with high chance of heart attack it was around 170BPM. Therefore, it is evident that a higher maximum heart rate is associated with a higher chance of heart attacks. Furthermore, the maximum heart rate was high for men than women out of those who were classified as having a higher chance of having a heart attack. (Figure-18b, appendix)

RQ3: As shown in Figure-19, out of those who have a higher chance of getting a heart attack, a vast majority of them have 0 major vessels coloured during the fluoroscopy test. Thus it is evident that those patients who had all 4 major arteries blocked were at the greatest risk of getting a heart attack

MODELS

Logistic Regression

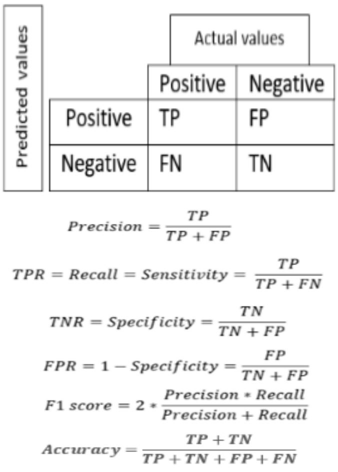
Logistic regression is a classification technique that uses a logit statistical model to predict the outcome of a binary dependent variable using a set of independent variables.

Assumptions

- Logistic regression is affected by multicollinearity. Thus, correlations between the independent variables should be assessed. According to (Tabachnick & Fidell, 2013), the assumption is met as long as correlations are less than 0.9.
- No outliers are present in the data
- Data is standardized
- Dependent variable is dichotomous in nature

Methodology

Five numerical features and eight categorical features were used as explanatory variables. The target variable was checked for skewness using a count-plot. As it was not heavily skewed (in the ratio of 1:100 or 1:1000) oversampling was not needed. A pipeline was created to impute, scale and one-hot encode the required features and thereafter the model was built.



Predicted values		Actual values	
		Positive	Negative
	Positive	TP	FP
	Negative	FN	TN

$$Precision = \frac{TP}{TP + FP}$$

$$TPR = Recall = Sensitivity = \frac{TP}{TP + FN}$$

$$TNR = Specificity = \frac{TN}{TN + FP}$$

$$FPR = 1 - Specificity = \frac{FP}{TN + FP}$$

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Figure 20

Analysis

A confusion matrix (Figure-20) evaluates a classification model by comparing the predictions and actual target values. It is used to measure accuracy, precision, recall and specificity. The 4 quadrants give: True Positives (TP), False Positives (FP -Type 1 error), True Negatives (TN) and False Negatives (FN - Type 2 error).

The Receiver Operator Characteristic (ROC) curve is a probability curve that plots the TPR vs FPR. The Area Under the Curve (AUC) measures the ability of the classifier to distinguish between the two classes. A higher AUC is desired.

When predicting heart attacks we want to be as accurate a possible but it is more of an issue when the model predicts low chance when in actuality there is a high chance as it is a matter of life or death. Thus, FNs should be reduced, hence the recall should be given high importance and aimed to be maximized. Therefore the AUC and recall will be given priority throught the analysis.

The model showed an accuracy and precision of 0.90 and a recall of 0.93 . The AUC was 0.94. There was no significant improvement from hyperparameter tuning so the initial model was selected as satisfactory as it had a very good recall and AUC.

Decision Tree Classifier

Please see ‘Decision Tree Regression’ under Task 2 for explanation of this model.

The algorithm operates as follows:

1. Pick the optimum attribute using attribute selection measures to split the data
2. Place the selected attribute as a decision node and break data into smaller subsets
3. Recursively repeat the above procedure for each child until: There are no more attributes, no more instances or all tuples are from the same attribute value.

Attribute selection methods (ASM): Information gain, Gini Ratio, Gini Index (Gini impurity).

Analysis

The initial model was distinctly improved upon by 3-5% on all scores after hyperparameter tuning. Gini impurity was selected as the best ASM to split the nodes with a max depth of four. The model showed a recall of 0.85 and an AUC of 0.90. Meaning out of everyone who has a high chance of heart attacks 85% were predicted positively. From a medical standpoint, 85% is not an adequate recall score for predictive models.

Gradient Boosting Classifier

Please see ‘XG Boost Regression’ under Task 2 for explanation of this model. Both XGBoost and this model follow the principle of gradient boosting. However, XGBoost uses a more regularized model formalization to prevent overfitting.

Analysis

The initial model showed slight improvements in accuracy, precision and AUC from hyperparameter tuning. However recall remained the same. Post tuning, we obtained a recall of 0.85 and an AUC of 0.93. Although the AUC was satisfactory, the recall score of this model was deemed subpar for practical use from a medical standpoint .

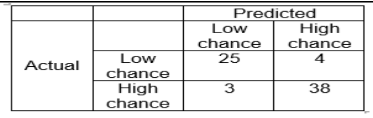
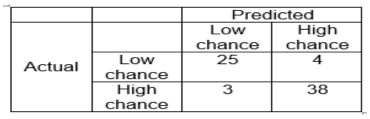
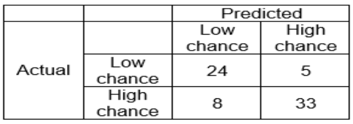
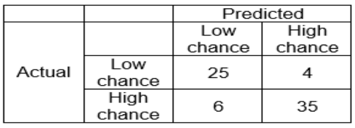
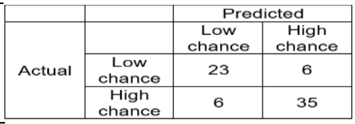
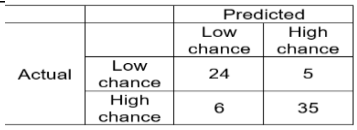
	Hyperparameters	Accuracy	Precision	Recall	F1-score	AUC	Confusion matrix
Logistic Regression	Before tuning	0.90	0.90	0.93	0.92	0.94	 Predicted Low chance High chance Actual Low chance 25 4 High chance 3 38
	After tuning	0.90	0.90	0.93	0.92	0.94	 Predicted Low chance High chance Actual Low chance 25 4 High chance 3 38
Decision Tree Classifier	Before tuning	0.81	0.87	0.80	0.84	0.82	 Predicted Low chance High chance Actual Low chance 24 5 High chance 8 33
	After tuning	0.86	0.90	0.85	0.88	0.90	 Predicted Low chance High chance Actual Low chance 25 4 High chance 6 35
Gradient Boosting	Before tuning	0.83	0.85	0.85	0.85	0.91	 Predicted Low chance High chance Actual Low chance 23 6 High chance 6 35
	After tuning	0.84	0.88	0.85	0.86	0.93	 Predicted Low chance High chance Actual Low chance 24 5 High chance 6 35

Figure 21

CONCLUSION

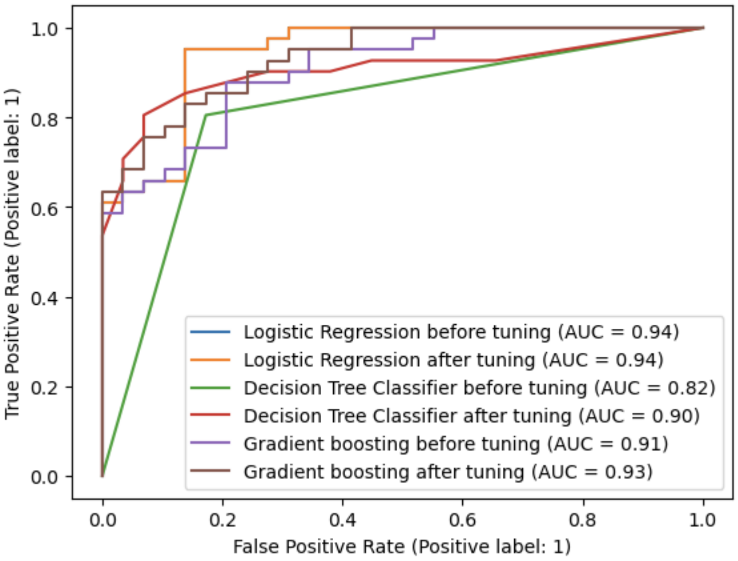


Figure 22

As this model is for medical grade purposes we aimed for a recall score and AUC of atleast 90%, such that the model minimizes the chance of predicting a heart attack imminent patient falsely and is also able distinguish between both cases as much as possible. Thus, the logistic regression model which showed a recall of 93% (Figure-21) and an AUC of 94% (Figure-22) was selected as the best model for predicting the chances of suffering a heart attack.

APPENDIX

BIBLIOGRAPHY

Abdullah, M.A.A. *et al.* (2016) *Annual wholesale distributor spending on products*, *Latest TOC RSS*. American Scientific Publishers. Available at: https://www.ingentaconnect.com/contentone/asp/asl/2016/00000022/00000012/art00020 (Accessed: April 1, 2023).

Fan, G.-Z., Ong, S.E. and Koh, H.C. (2006) *Determinants of house price: A decision tree approach*, *Determinants of House Price: A Decision Tree Approach*. Available at: https://journals.sagepub.com/doi/abs/10.1080/00420980600990928?journalCode=usja (Accessed: March 29, 2023).

Harrison, D. and Rubinfeld, D.L. (1978) *Hedonic housing prices and the demand for Clean Air*, *JOURNAL OF ENVIRONMENTAL ECONOMICS AND MANAGEMENT*. Available at: https://www.researchgate.net/profile/Daniel-Rubinfeld/publication/4974606_Hedonic_housing_prices_and_the_demand_for_clean_air/links/5c38ce85458515a4c71e3a64/Hedonic-housing-prices-and-the-demand-for-clean-air.pdf (Accessed: March 29, 2023).

Maas, A.H. and Appelman, Y.E. (2010) *Gender differences in coronary heart disease*, *Netherlands heart journal : monthly journal of the Netherlands Society of Cardiology and the Netherlands Heart Foundation*. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3018605/ (Accessed: April 3, 2023).

Mahmood, A.M. and Kuppa, M.R. (2010) *Early detection of clinical parameters in heart disease by improved Decision Tree Algorithm*, *2010 Second Vaagdevi International Conference on Information Technology for Real World Problems, Warangal, India, 2010*. Available at: https://ieeexplore.ieee.org/abstract/document/5692991/ (Accessed: April 2, 2023).

Perret-Guillaume, C., Joly, L. and Benetos, A. (2009) *Heart rate as a risk factor for cardiovascular disease*, *Progress in cardiovascular diseases*. U.S. National Library of Medicine. Available at: https://pubmed.ncbi.nlm.nih.gov/19615487/ (Accessed: April 3, 2023).

Popovic, R. (2009) *Effects of market structure changes on dairy supply chain in Serbia - IFAMA*. Available at: https://www.ifama.org/resources/files/2009-Symposium/1003_paper.pdf (Accessed: April 1, 2023).

Selim, S. (2008) *Determinants of house prices in Turkey: A hedonic regression model*, *TÜRKİYE'DE KONUT FİYATLARININ BELİRLEYİCİLERİ: HEDONİK REGRESYON MODELİ*. Available at: https://dergipark.org.tr/en/download/article-file/2151923 (Accessed: March 29, 2023).

LINKS FOR DATASETS

TASK 1: Wholesale customers- https://www.kaggle.com/datasets/binovi/wholesale-customers-data-set

TASK 2: Boston house prices- https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data

TASK 3: Heart attack analysis- https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

GLOSSARY OF TERMS AND ABBREVIATIONS

TASK 1

- **CHANNEL:** Customers Channel - 1 = Hotel (Hotel/Restaurant/Café HoReCa) or 2 = Retail
- **REGION:** Customers Region- 1 = Lisbon, 2 = Oporto, 3 = Other
- **FRESH:** annual spending (in monetary units/m.u) on fresh products
- **MILK:** annual spending (in monetary units/m.u) on milk products
- **GROCERY:** annual spending (in monetary units/m.u) on grocery products
- **FROZEN:** annual spending (in monetary units/m.u) on frozen products
- **DETERGENTS_PAPER:** annual spending (in monetary units/m.u) on detergents and paper products
- **DELICATESSEN:** annual spending (in monetary units/m.u) on and delicatessen products

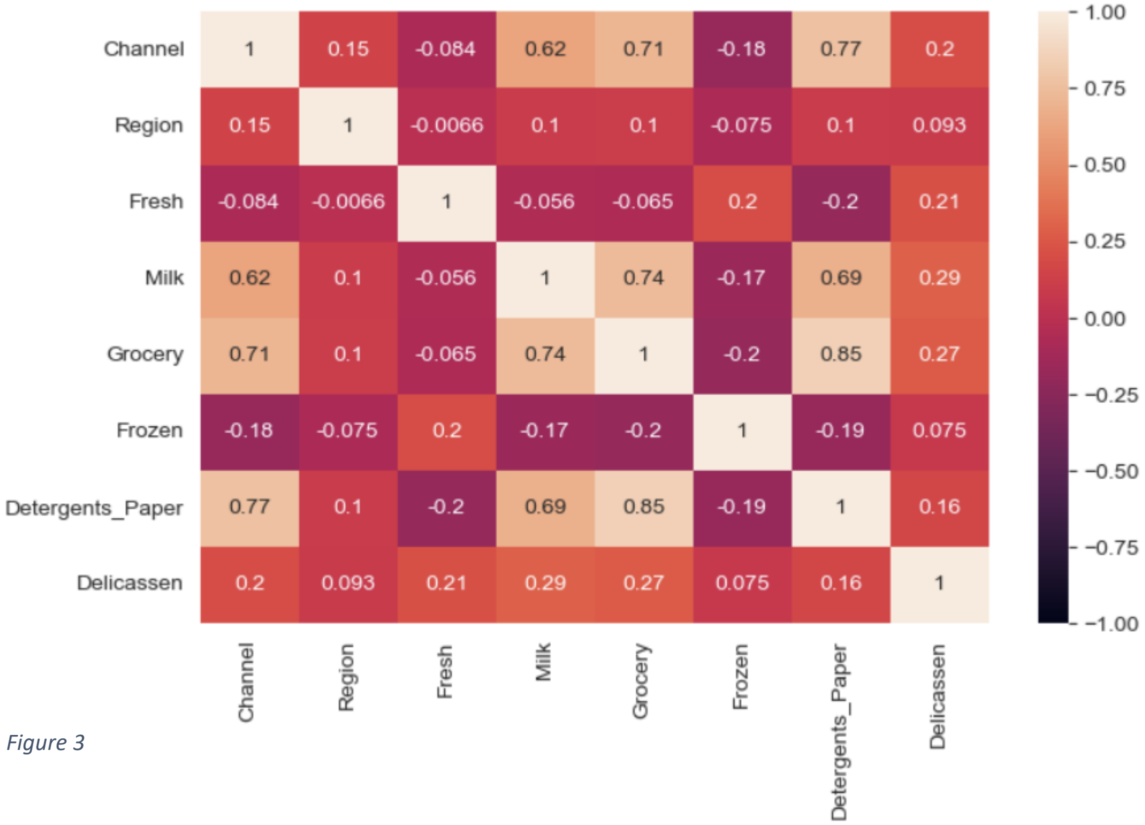
TASK 2

- **CRIM**: Per capita crime rate by town
- **ZN**: Proportion of residential land zoned for lots over 25,000 sq. ft
- **INDUS**: Proportion of non-retail business acres per town
- **CHAS**: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- **NOX**: Nitric oxide concentration (parts per 10 million)
- **RM**: Average number of rooms per dwelling
- **AGE**: Proportion of owner-occupied units built prior to 1940
- **DIS**: Weighted distances to five Boston employment centers
- **RAD**: Index of accessibility to radial highways
- **TAX**: Full-value property tax rate per $10,000
- **PTRATIO**: Pupil-teacher ratio by town
- **B**: $1000(Bk — 0.63)^2$, where Bk is the proportion of people of African American descent by town
- **LSTAT**: Percentage of lower status of the population
- **MEDV**: Median value of owner-occupied homes in $1000s

TASK 3

- **Age** : Age of the patient (in years)
- **Sex** : Gender (1 = male; 0 = female)
- **exng**: exercise induced angina (1 = yes; 0 = no)
- **caa**: number of major vessels (0-3) colored by fluoroscopy. Major cardial vessels are as goes: aorta, superior vena cava, inferior vena cava, pulmonary artery (oxygen-poor blood --> lungs), pulmonary veins (oxygen-rich blood --> heart), and coronary arteries (supplies blood to heart tissue)
- **cp** : Chest Pain type
  - Value 0: typical angina (all criteria present)
  - Value 1: atypical angina (two of three criteria satisfied)
  - Value 2: non-anginal pain (less than one criteria satisfied
  - Value 3: asymptomatic (none of the criteria are satisfied)
- **trtbps** : Resting blood pressure (in mmHg, upon admission to the hospital)
- **chol** : cholesterol in mg/dl fetched via BMI sensor
- **fbs** : fasting blood sugar > 120 mg/dL (likely to be diabetic) 1 = true; 0 = false
- **rest_ecg** : Resting electrocardiogram results --
  - Value 0: normal
  - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- **thalach** : maximum heart rate achieved (in BPM)
- **oldpeak**: ST depression induced by exercise relative to rest (in mm, achieved by subtracting the lowest ST segment points during exercise and rest)
- **slp**: the slope of the peak exercise ST segment, ST-T abnormalities are considered to be a crucial indicator for identifying presence of ischaemia –
  - Value 0: upsloping
  - Value 1: flat
  - Value 2: downsloping
- **thall** - Thalium Stress Test result (0,1,2,3)
- **output** : 0= less chance of heart attack 1= more chance of heart attack

TABLES AND FIGURES



	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Channel	1	0.15	-0.084	0.62	0.71	-0.18	0.77	0.2
Region	0.15	1	-0.0066	0.1	0.1	-0.075	0.1	0.093
Fresh	-0.084	-0.0066	1	-0.056	-0.065	0.2	-0.2	0.21
Milk	0.62	0.1	-0.056	1	0.74	-0.17	0.69	0.29
Grocery	0.71	0.1	-0.065	0.74	1	-0.2	0.85	0.27
Frozen	-0.18	-0.075	0.2	-0.17	-0.2	1	-0.19	0.075
Detergents_Paper	0.77	0.1	-0.2	0.69	0.85	-0.19	1	0.16
Delicassen	0.2	0.093	0.21	0.29	0.27	0.075	0.16	1

Figure 3

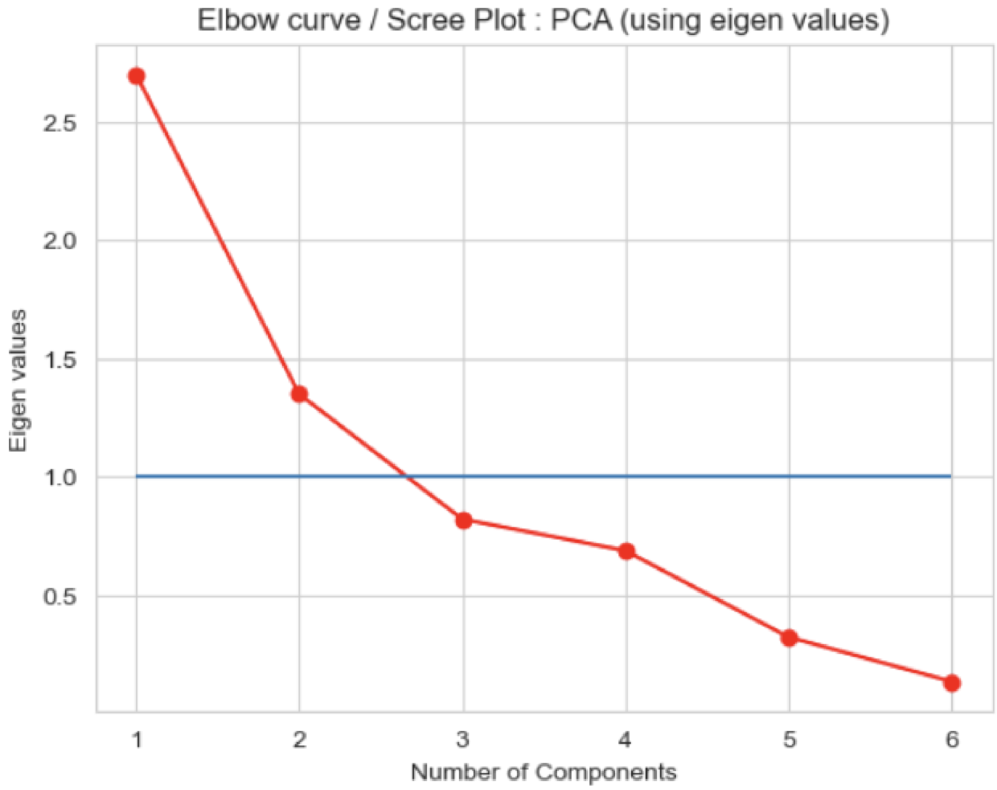


Figure 5b

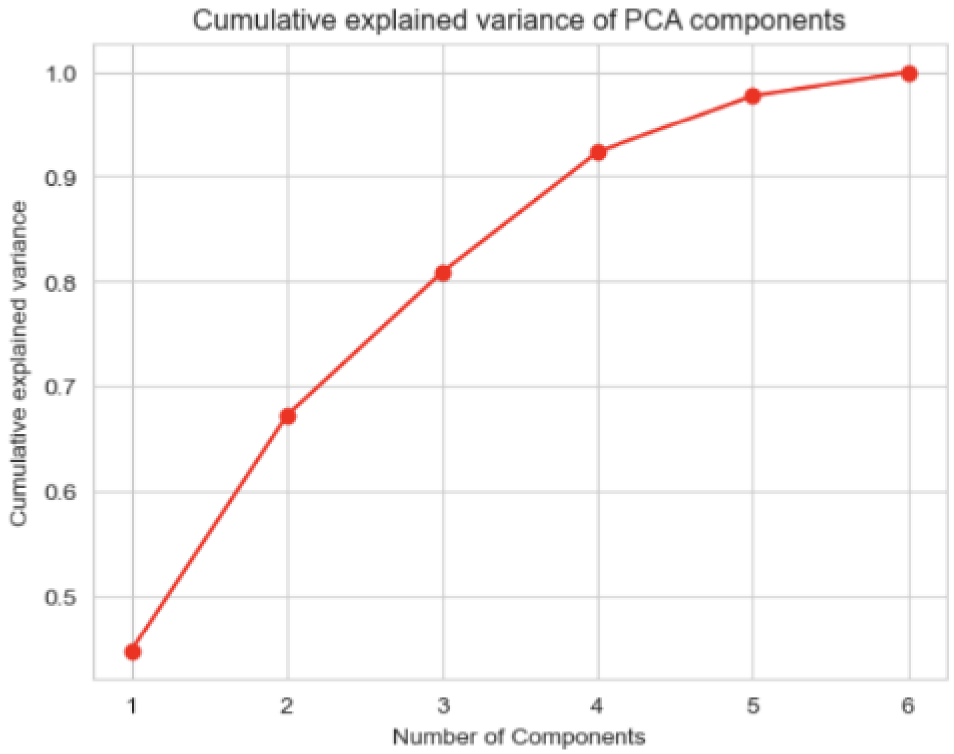


Figure 5c

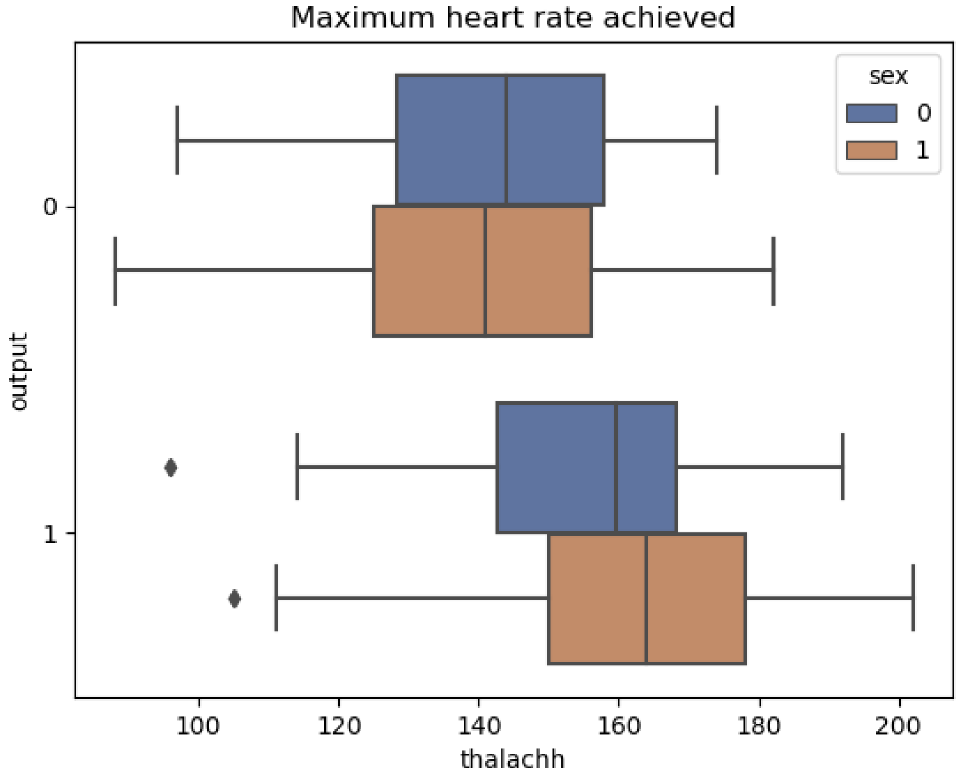


Figure 18b