MACHINE LEARNING

ST3189 Coursework

UNIVERSITY OF LONDON

UOL STUDENT ID: 210501867

Total 7 pages excluding cover page, table of content and references

Contents

Introduction

Machine Learning is the use of computer systems that can learn and modify on their own without specific instructions to reinforce on a specific task that you wish for it to do. It does so by using programming algorithms and statistic models to analyze patterns from the dataset given. In this coursework, we will only be looking at supervised and unsupervised learning. Supervised learning is the training of a statistical model to obtain a desired output value that is used together with input variables given in the dataset. Unsupervised learning is the training of a statistical model on unlabeled data where the algorithm learns and obtains patterns and/or relationships from it.

1. Movies Gross Performance – Regression Analysis

Knowing how much gross that might come from releasing a move is crucial for any production company to maintain profitability in this entertainment industry. Usually with high viewership comes with high revenue if the movie produced is good. However, to know exactly what drives its gross performance, production companies must analyze the data they have collected from viewership and reviews of the movie. Therefore, we will investigate how to effectively forecast a model to see how different variables affect their gross performance with the available dataset given. The following are research questions (RQ) that are identified:

- RQ 1: Which model shows the best result in predicting *gross* performance?
- RQ 2: Which variable had the most influence on *gross* performance?

1.1 Methodology

Use of Exploratory Data Analysis (EDA) to see any correlation between the independent variable and the target variables. We will then make use of Multiple Linear Regression, Random Forest, Support Vector Matrix and XGBoost that are regression techniques that help predict our dependent variable. Lastly, we will compare all the performance of models used to obtain the best fit model for prediction and find out what variable has the most influence on the gross performance.

1.2 Dataset

The movies performance dataset has a total of 7658 observations and 9 columns. No duplicates were found, and null values were mostly under the *budget*/*gross* columns, so we replaced them with 0 values.

Variables	**Description**
Name	Name of movie
Genre	Type of movie genre
Year	Year movie was produced
Score	IMBD score out of 10
Votes	Number of votes given by viewers
Budget	Movie budget
Gross	Gross made for the movie
Runtime	Duration of movie
Company	Production company

1.3 Research

After using EDA, we find out that certain categorical variables are not as important. They include *name*, *year,* and *company* as they have too many factors to consider. The only exception would be *genre*. We

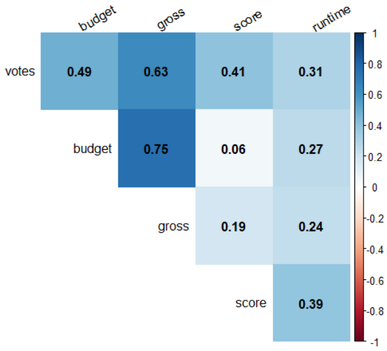


Figure 1. Correlation plot.

divide the dataset into test and trains sets at a 20:80 ratio before digging into our analysis. From the correlation matrix plot (Figure 1), we see that *gross* and *budget* have the strongest correlation with each other while *votes* come after. The categorical variables were removed as said previously that they did not have much significance after doing EDA except for *genre*. For the first model (linear regression) in multiple linear regression, we used all numerical variables and made dummy variables for *genre*. The second model (linear regression 2) had insignificant variables removed which included *runtime*, *score*, and multiple *genres* from the previous model to see if there were any differences to the model becoming significant. This in fact made the second model have a higher adjusted R-squared of 0.6618 compared to the previous model at 0.6615. In Figure 3, shows the residual plots of the two linear models. The red line is a smooth fit of the residuals, and it exhibits a slight curve at the start, showing some indications of non-linearity in the data. As the residual plots show non-linear associations in the data, we choose to use a non-linear transformation method using the I() function to help raise the independent numerical values to the power of 2 to change the linear relationship between variables. As seen in Figure 2, there is a slight improvement of how the variables are spread out along the residual = 0 line, with the residuals roughly forming a horizontal band around it. The following table shows the result of the three regression models we have made so far.

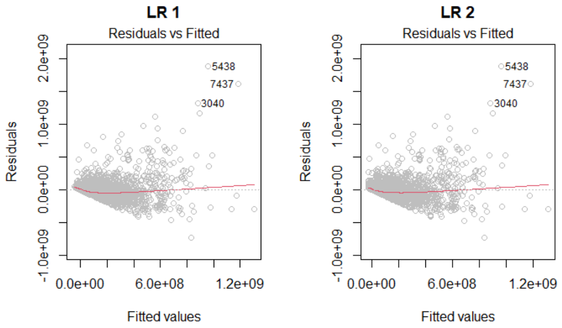


Figure 3. Correlation plot.

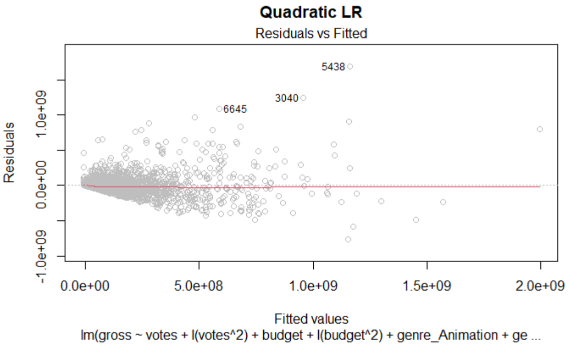


Figure 2. After non-linear transformation of predictors.

Models	**Adjusted R-Squared**	**RMSE**
Linear Regression 1	0.661584	87,368,985
Linear Regression 2	0.661801	87,368,985
Quadratic LR	0.706720	84,423,164

From the table above, we can observe that the Quadratic LR increased the overall adjusted R-squared from 0.6618 to 0.7067. This means a 4% increase of the variation of *gross* is explained by the independent variables used. It also has the lowest Root Mean Squared Error (RMSE) among the other 2 models. We will continue to analyze three other techniques: Random Forest (RF), Support Vector Machine (SVM) and Xtreme Gradient Boost (XGBoost).

Random Forest combines several decision trees that help strengthen predictive accuracy and reduce the possibility of overfitting. For our model, we used the default number of trees of 500. The R-squared and RMSE returned was 73.03% and 77,096,427 respectively. We also used the function, VarImp, to find out the most important variables from Random Forest. The top two most important variables are *budget* and *votes* seen in Figure 4.

Support Vector Machine objective is to find the best hyperplane with the largest margin between predicted values. It can be used for both regression and classification problems. In our case, 5918 support vectors were used but the RMSE returned was high at 162,413,205.

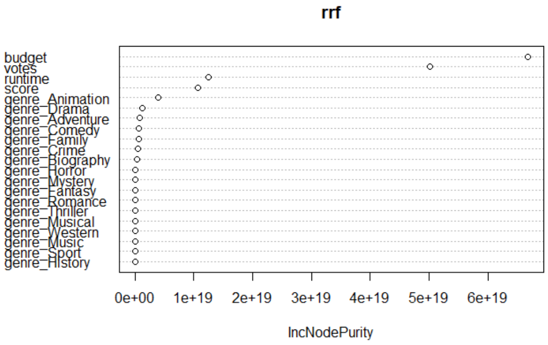


Figure 4. Important variables from RF.

XGBoost is a boosting algorithm. It takes the training data, repeatedly trains, and evaluate the model on testing data until the model stops improving. The RMSE returned is 81,609,432 with the top two most important variables being *budget* and *votes* as well.

1.4 Results

RQ 1: Which model shows the best result in predicting *gross* performance?

From our analysis and research, we can say that Random Forest is the best model to do the prediction, having the lowest RMSE of 77,072,280. This can be seen in Figure 5.

RQ 2: Which variable had the most influence on predicting *gross* performance?

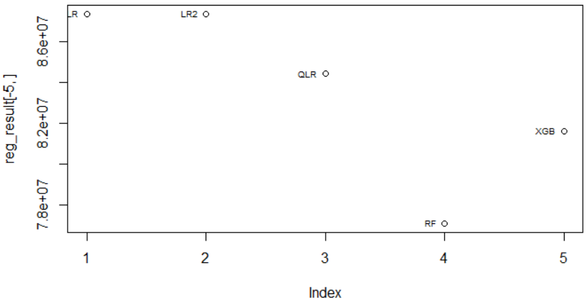


Figure 5. RMSE of all regression models.

From the results taken in Random Forest and XGBoost, the variables with the most influence are *budget* and *votes*.

2. Determining Heart Diseases – Classification Analysis

To be able to identify the issue or pain that a patient has is extremely crucial and timely for many hospitals worldwide. A survey was conducted in the US in 1977 where 21% of the adults responded that they had experienced medical error (Ihi.org, 2017). Though this percentage is still within acceptable range of 8% - 25% (Datacamp.com, 2023), it can be still considered a high error rate between that range. In this classification analysis, we will be focusing on identifying possible occurrence of heart diseases in patients through the different causes and/or pains that are given in our data. The following are research questions (RQ) that are identified:

- RQ 1: Which model shows the best result in predicting patients with heart diseases (*target)*?
- RQ 2: Which variable had the most influence on identifying *target*?

2.1 Methodology

Started off with some cleaning and mutation of data for easier reference. After which we went into EDA to observe any correlations between variables and how they react to one another in terms of categorical and continuous variables. We used Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine as our classification techniques. At the end, we will summarize our findings with the results of accuracy and area under curve (AUC) for the models made with the techniques mentioned.

2.2 Dataset

The heart dataset has a total of 1018 observations and 12 columns. No null values were seen and though we have 723 duplicate variables, it is due to the binary and categorical variables in the dataset.

Variables	Description
Age	Age of patient
Sex	Gender of patient
Cp	Chest Pain (0-3)
Trestbps	Resting blood pressure
Chol	Serum cholesterol in mg/dl
Fbs	Fast blood sugar
Restecg	Resting electrocardiographic results
Thalach	Maximum heart rate achieved
Exang	Exercise induced angina (1 = yes, 0 = no)
Ca	Number of major vessels (0-3)
Thal	Blood disorder (0-2)

2.3 Research

We started off with EDA after cleaning our imported datasets. We used two different categories, continuous and categorical variables. This will allow us to see correlations among the continuous variables as well as to pinpoint which categorical would be of importance to keep a look out for.

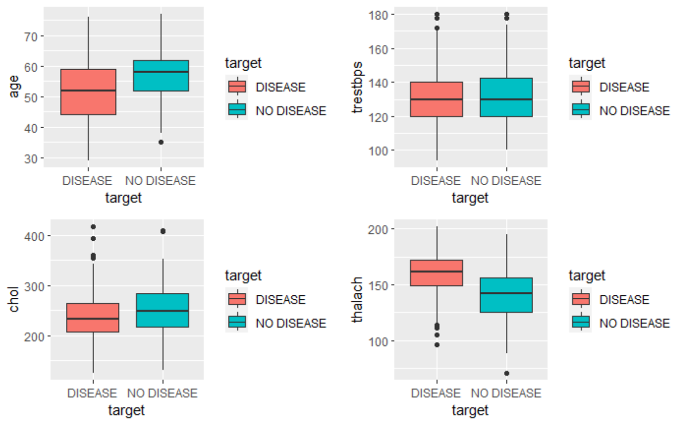


Figure 6. Boxplots for continuous variables.

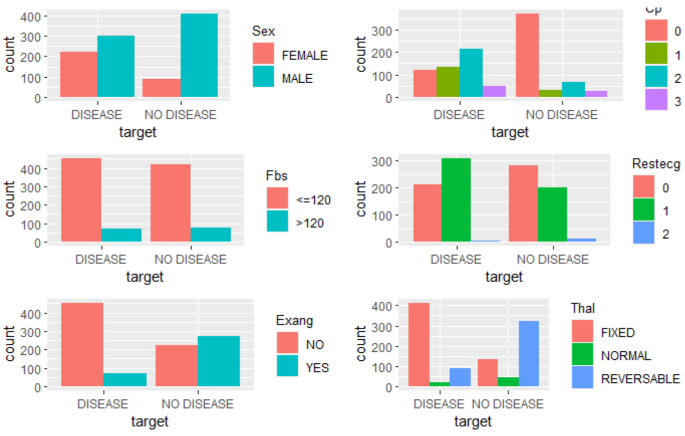


Figure 7. Bar plots for categorical variables.

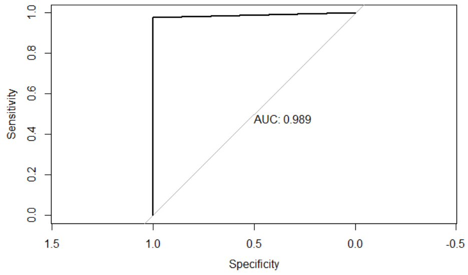
The correlation plot did not show any significant correlations between the continuous variables which could indicate that they require the mix of categorical variables for better analysis. We further analysed the continuous variables by using *target* as our main differentiator in our boxplots. It can be seen in Figure 6 that the age group of those who had heart disease was around 45 to 59. Solely using *trestbps* and *chol* will be difficult in determining patients with heart disease. Patients with *thalrach* more than 150 were seen to have heart disease.

As for the categorical variables in Figure 7, we find that females have a higher tendency to get a heart disease and that most patients with heart disease have chest pains on all levels. Having only *fbs*, *restecg* and *exang* to diagnose a patient with heart disease will be unreliable as their data comparison is similar or not making sense. Lastly, we can observe that most patients with heart disease have a fixed defect from *thal*.

We divided the dataset into 80:20 ratio for its training and test set which all three of our classification technique would be following for a more accurate comparison at the end. Starting with logistic regression (lgm), we used the glm function and set the family to “binomial” to inform the function that the dependent variable (*target*) is binary. Results showed that the AIC is at 610.10, where the lower the AIC the better the model. We then used the Variance Inflation Factor (VIF) to check on multicollinearity between variables and there was none. Only insignificant variables for lgm.

We then removed the two insignificant variables for the second model (lgm2) which were *fbs* and *age*. The AIC improved to 606.98 and VIF was clear of high values. All variables in lgm2 were significant hence we move on to making predictions.

Area Under Curve (AUC) shows how classification models performs with a range of 0.5 to 1 where 1 will be the perfect classification. We will also be using accuracy to finalise which model is the best. Using a threshold of 0.5, the accuracy taken from the confusion matrix shown 0.8346 while the AUC was at 0.8304.



Next, we used Random Forest model that used 500 trees with 3 variables tried at each split. The result of both accuracy and AUC was extremely high of 0.9882 and 0.98 respectively.

The Decision Tree model is a non-parametric supervised learning algorithm, which can be utilized for both classification and regression task (Psnet.ahrq.gov). The accuracy was 0.8583 and AUC at 0.8553.

Lastly, using of Support Vector Machine where it finds the best hyperplane that separates data into different classes. The result showed an accuracy of 0.8358 and AUC of 0.8336.

2.4 Results

RQ 1: Which model shows the best result in predicting patients with heart diseases (*target)*?

Models	**Accuracy**	**AUC**
Logistic Regression 2 (lgm2)	0.8346	0.8304
Random Forest	0.9882	0.9800
Decision Tree	0.8583	0.8553
Support Vector Machine	0.8358	0.8336

From the table above, we can conclude that Random Forest has the best model in predicting patients with heart diseases as its accuracy and AUC are the highest among the other models.

RQ 2: Which variable had the most influence on identifying *target*?

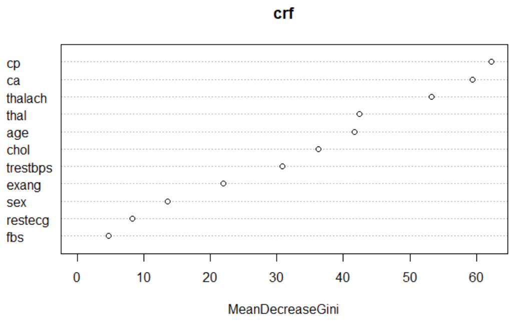


Figure 8. Important variables from RF for classification.

Taking results from the varImp function in our Random Forest model, *cp*, *ca* and *thalach* are the most important variables on identifying *target* correctly as seen in Figure 8.

3. Aeon Mall Customers – Unsupervised Learning

The mall has accumulated some of the customers data to help them target customers with the right products and price ranged advertisements. To analyse what sort of customer segments there are, we must determine the clusters and conclude what we have discovered from the unlabelled data. These will help us understand customers spending and for the marketing team to be able to plan their strategy accordingly. While doing unsupervised learning, we will discover new insights of the type of customers the mall has, and strategy needed to put in place. The following are research questions (RQ) that are identified:

- RQ 1: What is the optimal number of clustering for this data?
- RQ 2: Who are the target customers whom we can focus our marketing efforts on?

3.1 Methodology

We will be using Principal Component Analysis (PCA), K-means clustering and hierarchical clustering using Euclidean distance metric. These are unsupervised machine learning techniques to further understand the relationship between variables of the unlabeled data that helps to cluster them into significant groups. Elbow and Silhouette methods were used to obtain the optimal number of clusters to be used.

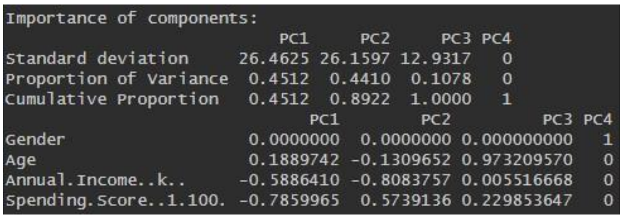
3.2 Dataset

The mall dataset has a total of 200 observations and 5 columns. No null or duplicate values were observed.

Variables	Description
CustomerID	Numbering the customers
Gender	Male or female customers
Age	Customer's age
Annual Income (k$)	Yearly income in thousands
Spending Score	Assigned by the mall based on customer behavior and spending nature

3.3 Research

We did EDA first to explore the types of customers the mall has. We mainly did histograms as it showed us the distribution of customers and the key information that is discovered. The results showed that the highest count of customers' age was between 30 to 35. Most of the customers' annual income is around $55,000 to $75,000 and their spending score, which was given by the mall, were equally distributed. With this information, we proceeded with K-means first followed by Hierarchical Clustering to discover more meaning behind this analysis. The first step was to do Principal Component Analysis (PCA). PCA is an unsupervised learning method that helps retain information on high-dimensional variables that has patterns which could be used to reduce complexity. We used PCA to find any sort of correlation between the numerical variables that could be seen in Figure 9. In the figure, we will only be using PC1 and PC2 because the remaining PCs' proportion of variance are close to 0.1 which suggests that it does not have that much interpretive value. In total, they account for 89.22% of the variance in the data. PCA on its own does not show us detailed relationships between the variables. Therefore, we are required to use K-means and Hierarchical Clustering to help us further understand it.



```
Importance of components:
                           PC1     PC2     PC3 PC4
Standard deviation     26.4625 26.1597 12.9317   0
Proportion of Variance  0.4512  0.4410  0.1078   0
Cumulative Proportion   0.4512  0.8922  1.0000   1
                              PC1        PC2         PC3 PC4
Gender                  0.0000000  0.0000000 0.000000000   1
Age                     0.1889742 -0.1309652 0.973209570   0
Annual.Income..k..     -0.5886410 -0.8083757 0.005516668   0
Spending.Score..1.100. -0.7859965  0.5739136 0.229853647   0
```

Figure 9. Principal Component Analysis.

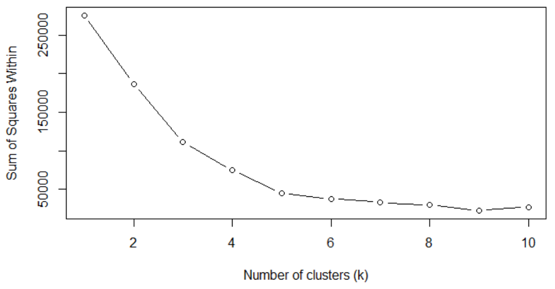


Figure 11. Elbow Method.

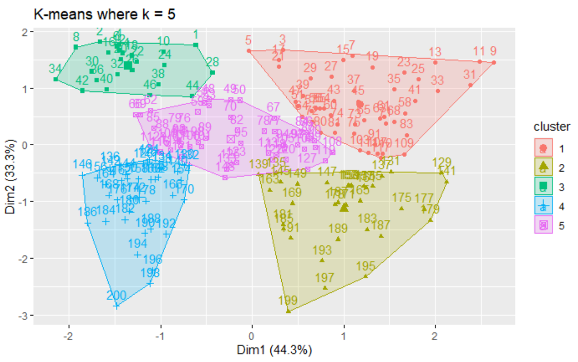


Figure 12. K-means Clustering.

The focus of K-means clustering is to separate n observations into k-clusters to the nearest mean. We first assume to use 3 clusters to calculate k-means and the result given within cluster sum of squares by cluster was 49.9%. After using the Elbow method (Figure 11) to find the optimal number of clusters, we could conclude that k = 5 was the best as it is observed that the bent starts there. The cluster sum of square by cluster improved was 75.6%, a 25.7% difference from the previous assumed cluster. In Figure 12, the green cluster represents customers with low annual income but high spending score. The blue cluster represents customers with low annual income and low spending scores. The pink cluster represents customers with middle income and middle spending scores. The red cluster represents customers with high income and high spending scores. The yellow cluster represents customers with high income but low spending scores.

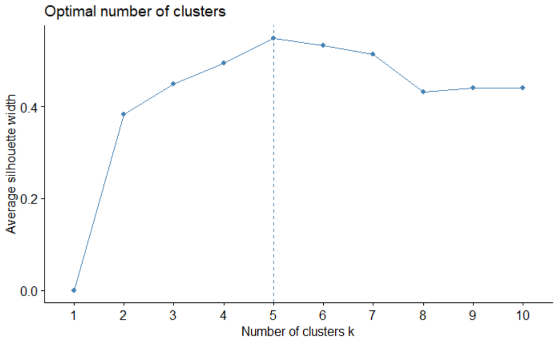


Figure 10. Silhouette Method.

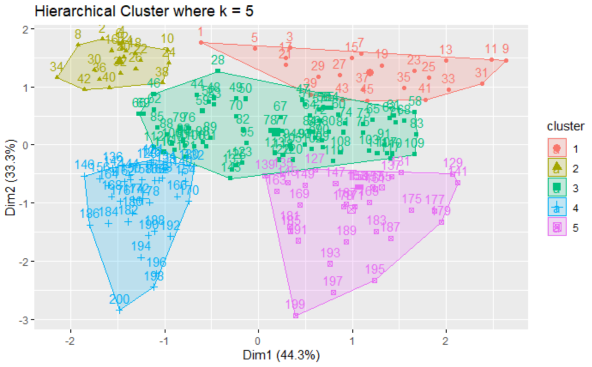


Figure 13. Hierarchical Clustering.

Hierarchical clustering is an algorithm that groups the same variables together that will then be called clusters. Each cluster are distinct from one another. From Figure 10, the silhouette method also showed the same result of k = 5 from the previous Elbow method plot. As for the clusters in Figure 13, they have similar groupings with K-means clustering.

3.4 Results

RQ 1: What is the optimal number of clustering for this data?

As seen from the Elbow and Silhouette method plots, the optimal number of clustering would be five.

RQ 2: Who are the target customers whom we can focus our marketing efforts on?

It depends on the aims and objectives of the mall. If we want to push on sales, we could market to the customers who have high income and/or high spending scores which can be identified through the clusters. The clusters involved could be yellow, green and red. If we want to raise the spending scores of customers, the marketing team can push promotional advertising to those who have low spending scores to entice them to spend at the mall. The clusters involved would be the blue and pink ones.

References

Americans' Experiences with Medical Errors and Views on Patient Safety. (2017). Institute for Healthcare Improvement. https://www.ihi.org/resources/publications/americans-experiences-medical-errors-and-views-patient-safety#:~:text=This%20report%20presents%20the%20findings,personally%20experienced%20a%20medical%20error

Decision Tree Classification in Python Tutorial. (2023, February). Learn Data Science and AI Online | DataCamp. https://www.datacamp.com/tutorial/decision-tree-classification-python

Medication Administration Errors. (2021, March 12). PSNet. https://psnet.ahrq.gov/primer/medication-administration-errors#:~:text=In%20a%20review%20of%2091,%E2%80%9325%25%20during%20medication%20administration