UOL Student Number: 200630447

Module: ST3189 (Machine Learning)

Number of Pages: 10

Machine Learning Report with Python

Table of Contents

1.0 Unsupervised Learning

1.1 Introduction

Unsupervised learning is called so, as an instructor is not available and there are no right or wrong responses. This method lets its algorithms understand some characteristics using the information provided and utilises these prior understood characteristics to identify the classes of new information, whenever they are provided. (Mahesh, 2020)

Task 1 makes use of the shop customer dataset which represents a thorough examination of an imaginative store's hypothetical clients, containing 2000 observations and 8 variables, such as customer ID, Age, Annual Income in $ and Profession and after importing the dataset, the data wrangling was done, resulting in 1976 observations and 7 variables left after cleaning.

1.2 Research Questions

Data was further visualised to identify certain correlations with each variable and through this, some research questions could be created and the graphs below show the correlations found using this dataset.

1. Does Annual Income differ according to profession?
2. Does Age differ with profession?
3. Does Gender affect annual income?

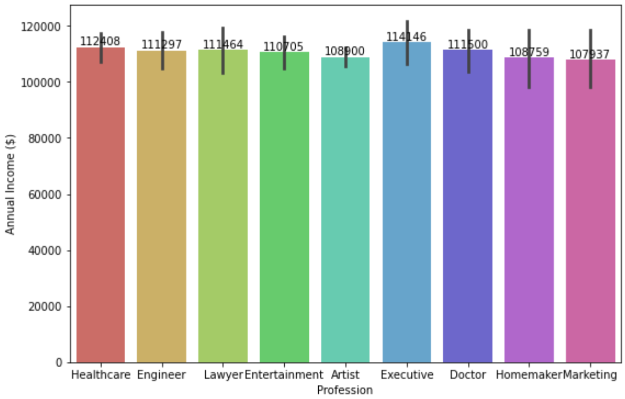

Figure 1: Annual Income levels for each profession

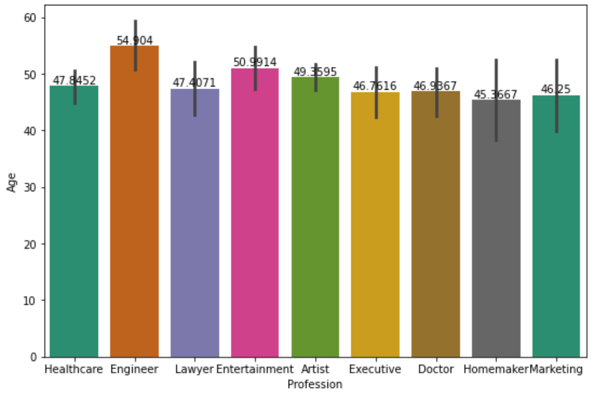

Figure 2: Age of workers for each profession

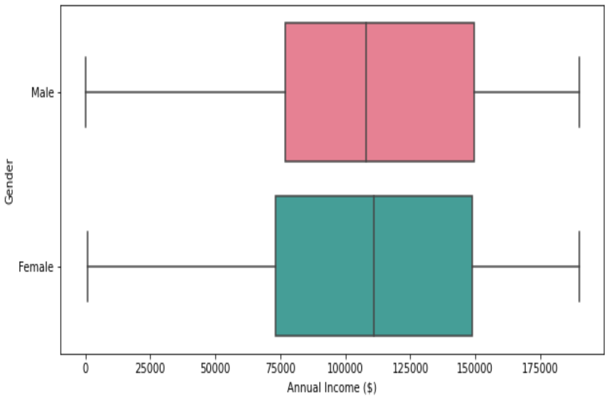

Figure 3: Annual Income for each gender

Figure 1 shows that executives earn the highest income, meaning the business professions tend to earn higher annual incomes. Figure 2 shows that engineers have the oldest average workers at around 55 years of age. Figure 3 illustrates that that males earn more than females when comparing annual incomes.

Literature for each conclusion

1) A study published on the article (IT, aircraft and marketing jobs are the highest paid in Macedonia, 2016) mentions that IT workers earn the highest salaries in Macedonia followed by aircraft workers and then marketing employees, which somewhat supports the conclusion drawn from the EDA which labels the business industry workers earning the higher incomes.

2) (Bosch & ter Weel, 2013) investigated the different jobs that older people had in the Netherlands from the years 1996 to 2010 and have concluded that they are usually employed in jobs that are declining and are considered lesser skilled, such as clerks and bookkeepers, which contradicts the assumption derived from the EDA of the dataset, stating engineers have older workers.

3) A study by (Hong Vo, Van, Tran, Vu, & Ho, 2019) found out a distinct gap in wages for men and women in Vietnam through results of investigation from years 2004 to 2016, where the gender income inequality favours the male gender. This supports the conclusion made from the EDA using the shop customer dataset stating the males out earn females.

1.3 Principal Component Analysis (PCA)

PCA is used as a dimension reduction technique for sizable datasets by making interpretation easier and limiting the loss of data through the creation of a new group of variables that enhance variance. (Jolliffe & Cadima, 2016).

The categorical variables were encoded through LabelEncoder and all the variables were standardised to make sure they contribute equally to the models to avoid bias. Next, a graph was used to find the Explained Variance Ratio, which shows how effective the principal components are and how many to choose for the model and shows the variance percentage that is linked to each of the principal components. (Lindgren, 2020). The ratio for this dataset was illustrated on Figure 4, showing that selecting 6 components would help target around 85% of the explained variance and this number is good as being over 80% helps to prevent overfitting.

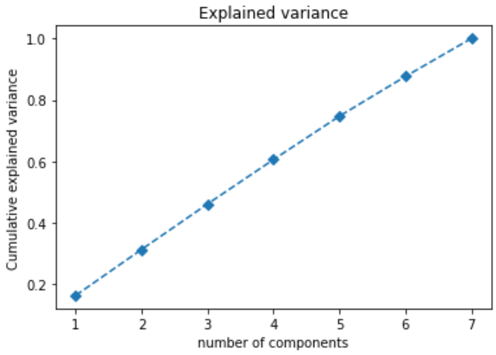

Figure 4: Explained variance graph

.

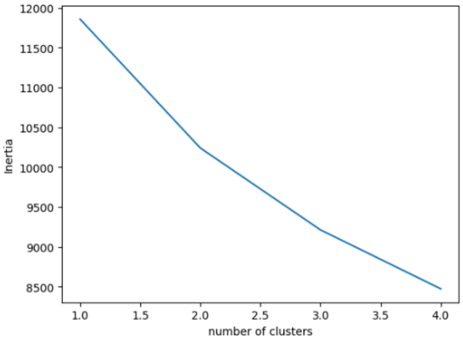

Figure 5: Elbow Graph(PCA)

Elbow graphs, as seen in Figure 5 show the best number of clusters to choose for K-means clustering, and based on the figure, 3 is the best number of clusters for K-means clustering, using PCA for this dataset, as the slope goes down into a continuous line from there on. Based on Figure 6, it can be seen that clusters 0 and 1 have higher income customers, earning until around $155,000 annually, while cluster 2 customers earn until around $125,000 annually and Figure 7 shows that clusters 0 and 2 have the lower Spending Score customers compared to cluster 1.

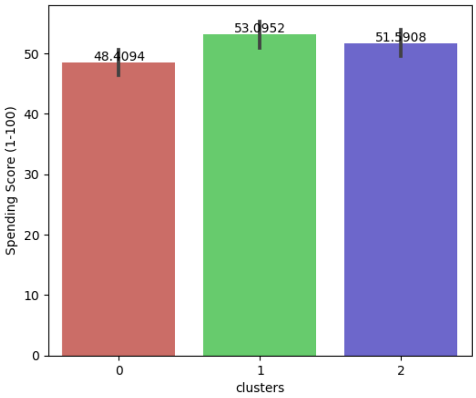

Figure 7: Spending score for each cluster

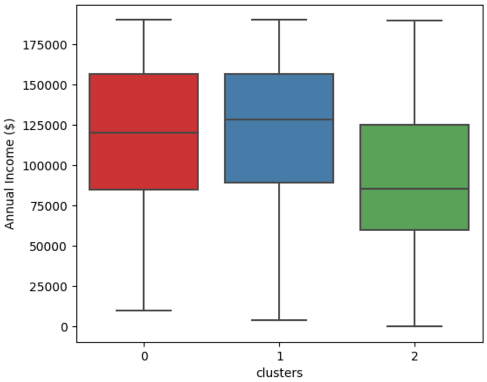

Figure 6: Annual Incomes for each cluster

1.4 K-Means

(Li & Wu, 2012) defines K-Means as a clustering/unsupervised algorithm, that is established on dividing, used in recognising configurations and mining data and is known to be brief and efficient.

The same visualisations were done for K-means clustering without PCA, where the number of clusters that is the best was still 3, using the elbow graph illustrated in Figure 8. Further, the scatter plot showed discernible clusters as seen in Figure 9, where it could be said cluster 1 (blue) showed annual income up to around $80,000, cluster 2 (yellow) described up to $130,000 and cluster 3 (pink) showed upwards of $175,000.

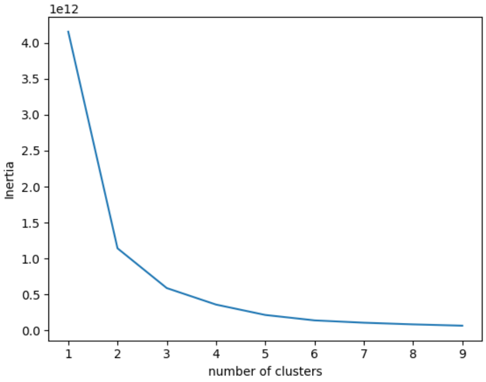

Figure 8: Elbow Graph(No PCA)

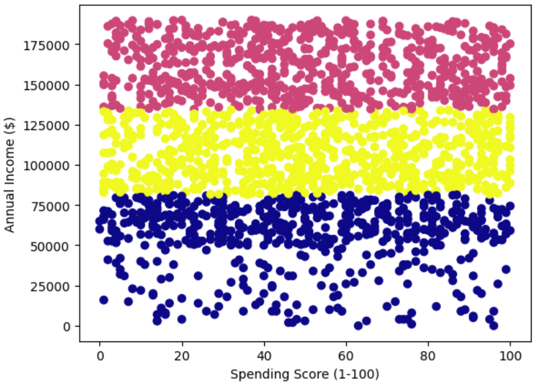

Figure 9: Scatter plot for each cluster with Spending score against Annual Income

2.0 Regression

2.1 Introduction

(Maulud & Abdulazeez, 2020) describes regression as a method for two uses: one for predictions and extrapolations and two, for finding the causes for the relationships between the dependent and predictor variables.

Task 2 made use of a used car dataset highlighting the brands of used cars including the kilometres they have driven, the gear transmission, the fuel type and many more, out of which it would be possible to create a regression model to predict prices. The dataset consisted of 4340 data values with 8 columns, and after cleaning it was left with 3218 data values at 7 columns (name was used for visualisation only).

2.2 Research Questions

As done with the unsupervised dataset, the regression dataset variables were also visualised to find correlations with selling price, the target variable and some research questions were created along with their graphs as shown below:

1. Does Fuel Type affect selling price?
2. Does the Year the car was bought affect selling price?
3. Does the Gear Transmission of the car affect selling price?

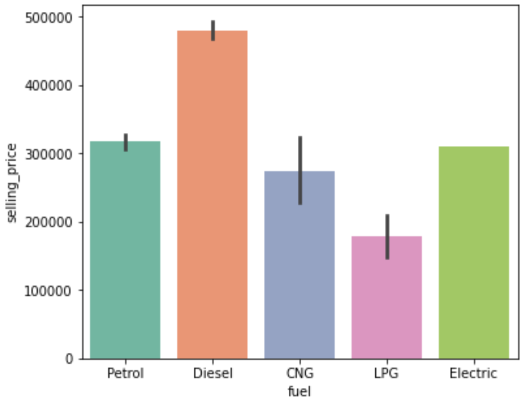

Figure 10: Most expensive fuel types

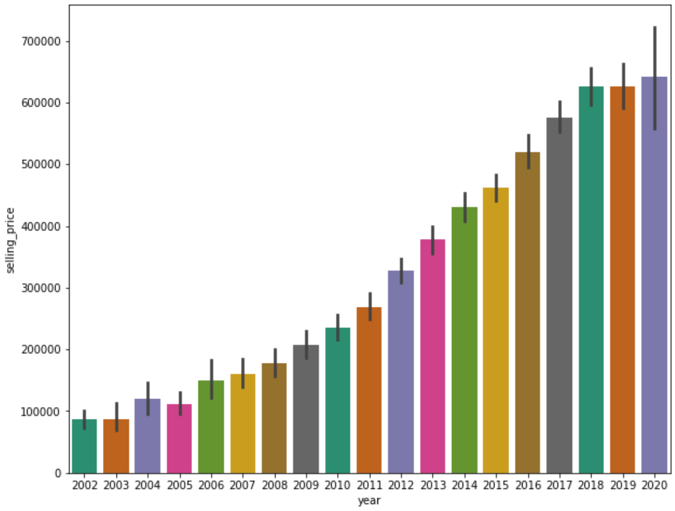

Figure 11: Price range over the years

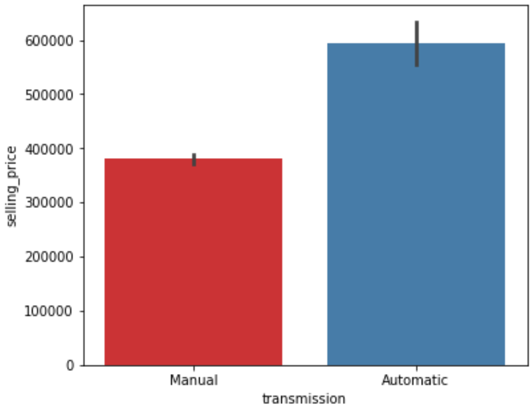

Figure 12: Most expensive transmission type

Figure 10 shows off that diesel vehicles have a higher selling price in comparison to the others. Figure 11 illustrates how selling price gradually increases as the years go on. Figure 12 shows that automatic transmission vehicles are more expensive than manuals

Literature for each conclusion

1) (Lebeau, Lebeau, Macharis, & Mierlo, 2013) mention that electric vehicles are more efficient when concerning energy, but they have found out that these vehicles usually are more expensive with compared to diesel and petrol automobiles, which contradicts the assumption derived from Figure 10, stating diesel cars are the most expensive fuel type.

2) (Cantillo & Ortúzar, 2014) studied the number of vehicles on the road and congestions, which could create many externalities and indicated that a lot of brand-new vehicles are bought but older vehicles are bought as well at a cheaper price for those with lesser incomes, supporting the conclusion shown in Figure 11 that older cars indeed have lower selling prices.

3) According to (DiLullo, Kocienski, & Zopatti, 2013), there was a sudden increase in the number of manual transmission cars to 7% of car sales and one of the reasons was said to be the lower purchase prices and better mileage. This supports the conclusion made from Figure 12 stating the automatic car prices are higher than manual transmission ones.

1.3 Regressor Models

- Multivariate Linear Regression is where the relation between various predictor variables is defined in terms with the dependent variable. (Kalogeropoulos, 2022)
- Random Forests creates multiple decision trees which are individual regression functions and concludes by giving the average of all the tree's outputs. (Li et al., 2018)
- Boosting is used to transform lazy leaners to eager learners and is used to reduce variance and bias. (Mahesh, 2020). The gradient boosting models used here were LightGBM and XGBoost, where both are established off decision trees and create asymmetric trees, but LightGBM grows the trees leaf wise, while XGBoost grows the trees level wise.

Through these models, the R-squared, Mean Absolute Error (MAE), Root Mean Squared Error(RMSE) and Mean Squared Error(MSE) were found. R-squared shows how much of the variance of the dependent variable is explained by the model, RMSE is "is a quadratic measure of the error between predicted and observed values" and MAE shows "the absolute error between predicted and observed values." (Regression Performance, 2021).

Based on these four evaluators, the better model depends on how high its R-squared score is and how low its MAE, MSE and RMSE are.

The values for each of the performance evaluators are shown below in Table 1:

Models	R-squared	MAE	MSE	RMSE
Linear Regression	0.534	131436	29077643295	362.54
RandomForest	0.509	131709	30653542812	362.92
XGBoost	0.537	125871	28897871632	354.78
LightGBM	0.571	121307	26787890295	348.29

Table 1: Performance evaluators before tuning

Based on these scores, it could be said that LightGBM fits the model better, as it has the highest R-squared score and yields lower MAE, MSE and RMSE, in comparison to the three other models used.

Linear Regression does not consist of any hyperparameters that can be adjusted, so after tuning the parameters for the other models, the scores for each have changed as shown below:

Models	R-squared	MAE	MSE	RMSE
RandomForest	0.591	119124	25537063680	345.14
XGBoost	0.599	118101	25031062352	343.66
LightGBM	0.598	118353	25066095355	344.02

Table 2: Performance evaluators after tuning

XGBoost now has the highest R-squared score when compared to the others and has the lower MAE, MSE AND RMSE in comparison, so XGBoost is now the best fitting model for this dataset following tuning.

Based on feature importance through the Decision Tree Regressor, it could be seen on the right, that the year had the highest score, showing that it had the largest impact on the selected models when it came to predictions.

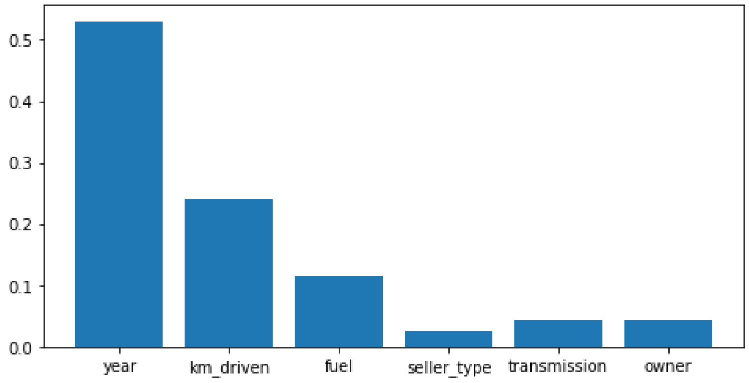

Figure 13: Feature Importance scores

3.0 Classification

3.1 Introduction

Classification is a method where the input group of data is categorised into classes depending on the variables given. It is part of supervised learning. (Joshi, 2022)

Task 3 used the Pima Indians Diabetes dataset, which consists of independent variables such as Age, BMI, Glucose and more, while the dependent variable was Outcome, where 0 was that the person did not have diabetes and 1 that the person had diabetes. The dataset had 768 observations and 9 variables and after data cleaning, it was left with 549 observations.

3.2 Research Questions

This dataset was visualised by comparing the predictor variables to the dependent variable outcome, and the following research questions and graphs were derived:

1. Does Age affect chances of diabetes?
2. Does Body Mass Index (BMI) affect chances of diabetes?
3. Does Blood Pressure affect chances of diabetes?

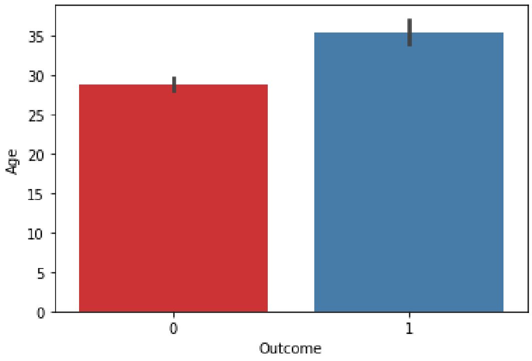

Figure 14: Outcome against Age

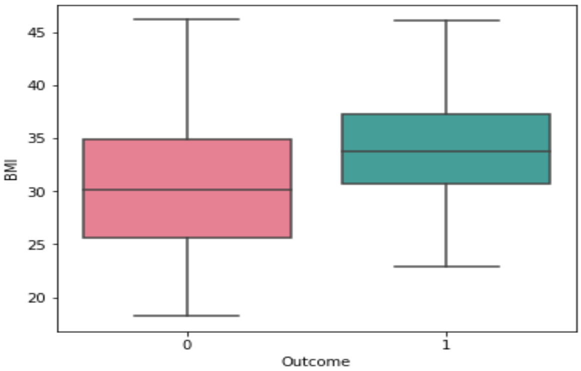

Figure 15: Outcome against BMI

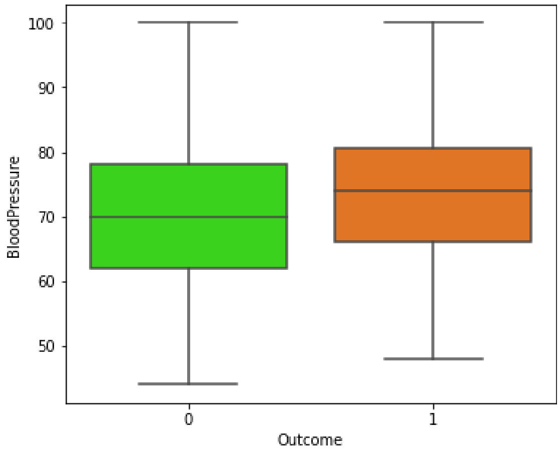

Figure 16: Outcome against Blood Pressure

Figure 14 shows that the lower the age, the lesser chances of diabetes. Figure 15 illustrates that lower BMI reduces the risks of diabetes. Figure 16 highlights that high blood pressure might increase chances of diabetes.

Literature for each conclusion

1) (Zoungas et al., 2014) highlighted the fact that there is a continuous increase in diabetes patients consisting of the age group of 20 to 40 years, and according to the 2013 report by International Diabetes Federation, more than 5 million of the deaths caused by diabetes were of people under 60 years of age, which contradicts with the conclusion that states that younger people have lower chances of diabetes.

2) (Gray, Picone, Sloan, & Yashkin, 2015) tried to see the consequences of a higher BMI on type 2 diabetes mellitus (DM) for the older U.S. citizens and concluded that higher BMIs do indeed cause continuously elevated risks of developing DM difficulties, which supports the conclusion illustrated by Figure 15 that higher BMI does increase chances of diabetes.

3) An article written by (Tziomalos & Athyros, 2015) identified the significant risk aspects that would cause and worsen diabetic nephropathy, which is a major reason for renal disease's end-stage. Some of the factors they have highlighted were obesity and high blood pressure, which supports the conclusion shown off by Figure 16, stating that higher blood pressure increases chances for testing positive for diabetes.

3.3 Models and Performance Evaluators

The models used here include decision tree classifier, logistic regression and the boosting methods of XGboost, LightGBM and Catboost.

- Decision Tree classifiers conduct classification of multiple classes on a dataset and predicts the smallest index class in the situation of many similarly high probability classes. (1.10 Decision Trees, 2023)
- Logistic Regression is another classification technique which is used when predicting binary events; in this dataset, that event is the occurrences of Outcome being either 0 or 1, using the predictor variables.
- Catboost is another gradient boosting method but compared to the other two boosting methods, it creates symmetric trees.
- Precision calculates how close the results from the model are to each other, while accuracy shows how near the results are to the measurement's definite value. (Hendricks, n.d.)
- Recall shows how well the model identifies accurate positive predictions based on the aggregate positive predictions.

- F1 score merges Recall and Precision to see the percentage of how well a model predicts accurately.
- Confusion matrices assesses a model's accuracy to distinguish predicted and actual classes. In the matrix, upper left indicates True Positive (TP), upper right is False Positive (FP), lower left is False Negative (FN) and lower right is True Negative (TN).

When splitting the dataset into training and testing datasets according to an 80 to 20 ratio, an additional step was taken, as the samples taken for Outcome being 0 and Outcome being 1 were imbalanced, being 393 (71.6%) and 156 (28.4%) respectively, as shown below, therefore, using oversampling through the resample technique, both Outcomes were made equal, to avoid any bias and improve the models' performance.

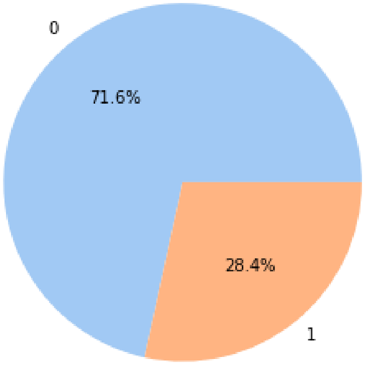

Figure 17: Pie chart illustrating percentage of samples for each Outcome; 0 is healthy and 1 is diabetic

The models were then fitted and the following accuracies, precision, F1 scores and confusion matrices were obtained for each, as show below:

Model	Accuracy	Precision	F1 Score	Confusion Matrix	
Decision Tree	70.00%	64.74%	65.16%	TP: 59	FP: 19
				FN: 14	TN: 18
Logistic Regression	80.00%	76.54%	77.60%	TP: 62	FP: 16
				FN: 6	TN: 26
XGboost	81.82%	77.96%	77.96%	TP: 68	FP: 10
				FN: 10	TN: 22
LightGBM	80.91%	77.12%	78.16%	TP: 64	FP: 14
				FN: 7	TN: 25
Catboost	80.00%	76.02%	76.95%	TP: 64	FP: 14
				FN: 8	TN: 24

Table 3: Performance evaluators before tuning

Based on this, it can be seen that XGboost has the highest accuracy and precision when compared to the other four models and further has the percentage of True Positive values. However, LightGBM has the highest F1 score and Logistic Regression has the highest percentage of True Negative values, according to the confusion matrices, but since accuracy, precision and True Positive percentage is still higher with XGboost, it is the best fitting model for this dataset.

To check if accuracies, precision and F1 scores could go higher, the parameters were tuned further and the following evaluators were obtained:

Model	Accuracy	Precision	F1 Score	Confusion Matrix	
Decision Tree	70.91%	67.42%	67.84%	TP: 56	FP: 22
				FN: 10	TN: 22
XGboost	83.64%	80.18%	81.42%	TP: 65	FP: 13
				FN: 5	TN: 27
LightGBM	81.82%	77.93%	78.72%	TP: 66	FP: 12
				FN: 8	TN: 24
Catboost	86.36%	83.14%	83.90%	TP: 69	FP: 9
				FN: 6	TN: 26

Table 4: Performance evaluators after tuning

Hyperparameter tuning for Logistic Regression was not done, as it does not have any significant parameters to influence accuracy increases. It can be seen above that Catboost has the highest accuracy, precision and F1 score when compared to the other models. Further, Catboost has the highest True Positive percentage but XGBoost has the highest True Negative percentage. But as mentioned above, since Catboost has both the higher precision, F1 score and True Positive percentage, Catboost is the now determined to be the best fitting model for this dataset.

The Receiver Operating Characteristic Curve (ROC) was also visualised for the models, shown in Figure 18, which illustrates the performance of each classifier against each other, Decision Tree has the highest Area Under the Curve of 0.96, therefore it can be said that Decision Tree is the best model for this dataset when it comes to differentiating negative classes with positive classes.

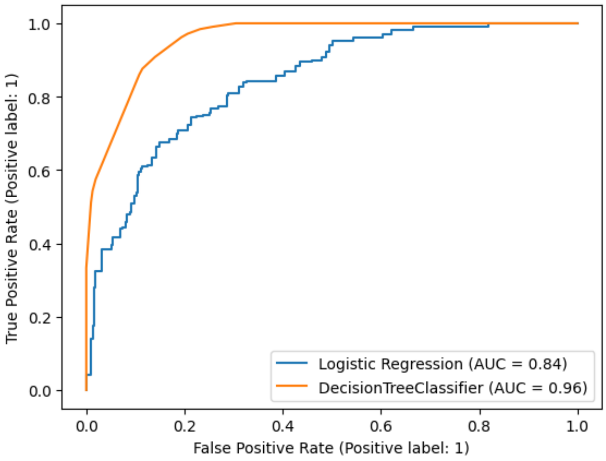

Figure 18: ROC curves

4.0 Bibliography

Note: The Regression Performance article had no author, therefore the in-text citation was labelled as the article name instead, same as the IT workers journal. Further, the study guide was also referenced with the in-text citation being (Kalogeropoulos, 2022) and the website What is the difference between accuracy and precision has no publishing date.

- *1.10 Decision Trees*. (2023). Retrieved March 23, 2023, from scikit-learn: https://scikit-learn.org/stable/modules/tree.html#:~:text=DecisionTreeClassifier%20is%20a%20class%20capable,class%20classification%20on%20a%20dataset.&text=In%20case%20that%20there%20are,lowest%20index%20amongst%20those%20classes.

- Bosch, N., & Ter Weel, B. (2013). Labour-market outcomes of older workers in the Netherlands: Measuring job prospects using the Occupational Age Structure. *De Economist, 161*(2), 199-218. doi:10.1007/s10645-013-9202-8
- Cantillo, V., & Ortúzar, J. D. (2014). Restricting the use of cars by license plate numbers: A Misguided Urban Transport Policy. *DYNA, 81*(188), 75-82. doi:10.15446/dyna.v81n188.40081
- DiLullo, G., Kocienski , S., & Zopatti, D. (2013, April 25). Development of Zero-Leg Input Manual Transmission Driving Interface. 99. Retrieved March 23, 2023, from https://web.wpi.edu/Pubs/E-project/Available/E-project-042413-122016/unrestricted/Assistance_Driving_Device_Project_Report.pdf

- Gray, N., Ph.D, Picone, G., Ph.D, Sloan, F., Ph.D, & Yashkin, A., Ph.D. (2015). The Relationship between BMI and Onset of Diabetes Mellitus and its Complications. *Southern Medical Journal, 108*(1), 29-36. doi:10.14423/SMJ.0000000000000214
- Hendricks, R. (n.d.). *What is the difference between precision and accuracy?* Retrieved March 23, 2023, from Deepchecks: https://deepchecks.com/question/what-is-the-difference-between-precision-and-accuracy/#:~:text=Accuracy%20vs%20Precision%20in%20Machine%20Learning&text=Machine%20Learning%20precision%20measures%20how,analogy%20to%20demonstrate%20their%20distinction.

- Hong Vo, D., Van, L. T., Tran, D. B., Vu, T. N., & Ho, C. M. (2019). The determinants of gender income inequality in Vietnam: A Longitudinal Data Analysis. *Emerging Markets Finance and Trade, 57*(1), 198-222. doi:10.1080/1540496x.2019.1609443
- IT, aircraft and marketing jobs are the highest paid in Macedonia. (2016, April). *Macedonian Business Monthly, 16*(163). Retrieved March 22, 2023, from https://go.gale.com/ps/i.do?p=STND&u=ull_ttda&id=GALE|A455405745&v=2.1&it=r&sid=summon

- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202. doi:10.1098/rsta.2015.0202
- Joshi, K. (2022, November 23). What is classification in Machine Learning and Why is it important? Retrieved March 23, 2023, from https://emeritus.org/blog/artificial-intelligence-and-machine-learning-classification-in-machine-learning/
- Kalogeropoulos, K. (2022). *Machine Learning.* London: University of London. Retrieved March 23, 2023
- Lebeau, K., Lebeau, P., Macharis, C., & Mierlo, J. V. (2013). How expensive are electric vehicles? A total cost of ownership analysis. *2013 World Electric Vehicle Symposium and Exhibition (EVS27)*. doi:10.1109/evs.2013.6914972
- Li, Y., & Wu, H. (2012). A clustering method based on K-means algorithm. *Physics Procedia, 25*, 1104-1109. doi:10.1016/j.phpro.2012.03.206
- Li, Y., Zou, C., Berecibar, M., Nanini-Maury, E., Chan, J. C., Van den Bossche, P., . . . Omar, N. (2018). Random Forest regression for online capacity estimation of lithium-ion batteries. *Applied Energy, 232*, 197-210. doi:10.1016/j.apenergy.2018.09.182
- Lindgren, I. (2020, April 24). Dealing with highly dimensional data using principal component analysis (PCA). Retrieved March 24, 2023, from https://towardsdatascience.com/dealing-with-highly-dimensional-data-using-principal-component-analysis-pca-fea1ca817fe6
- Mahesh, B. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJSR), 9*(1). doi:10.21275/ART20203995
- Maulud, D. H., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends, 1*(4), 140-147. doi:10.38094/jastt1457
- *Regression performance*. (2021, April 21). Retrieved March 23, 2023, from C3 AI: https://c3.ai/introduction-what-is-machine-learning/regression-performance/
- Tziomalos, K., & Athyros, V. G. (2015). Diabetic nephropathy: New risk factors and improvements in diagnosis. *The Review of Diabetic Studies, 12*(1-2), 110-118. doi:10.1900/RDS.2015.12.110
- Zoungas, S., Woodward, M., Li, Q., Cooper, M. E., Hamet, P., Harrap, S., . . . Chalmers, J. (2014). Impact of age, age at diagnosis and duration of diabetes on the risk of macrovascular and microvascular complications and death in type 2 diabetes. *Diabetologia, 57*, 2465-2474. doi:10.1007/s00125-014-3369-7