Module: Machine Learning (ST3189)

UoL Student Number: 210457922

Page Count: 10 pages (Excluding Cover Page, Table of Contents and Bibliography)

Contents

Task 1 - Unsupervised learning

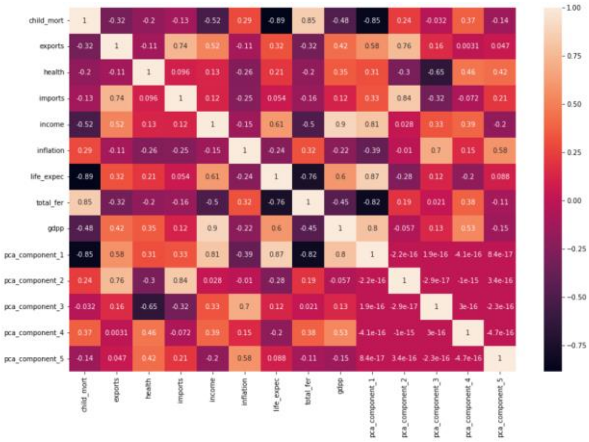
Unsupervised learning, uses machine learning algorithms to analyze and cluster unlabeled datasets (IBM). These algorithms enable us to discover patterns within the data and categorise them into homogenous groups. Additionally, it's used for dimension reduction which streamlines the data and improves the accuracy of the model by extracting the irrelevant variables and retaining only the crucial ones.

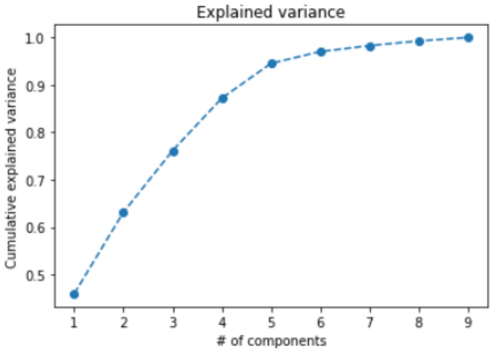
The dataset used for this task contains socio-economic and health data about 167 countries, including information such as the inflation rate, GNI, child mortality rate and 7 other variables. We aim to carry out a dimensionality reduction technique known as PCA followed by K-Means clustering to group countries into homogeneous clusters that can then be used to determine the overall development of the country in order to aid decision makers on how to prioritize funds towards countries that need it the most.

PCA

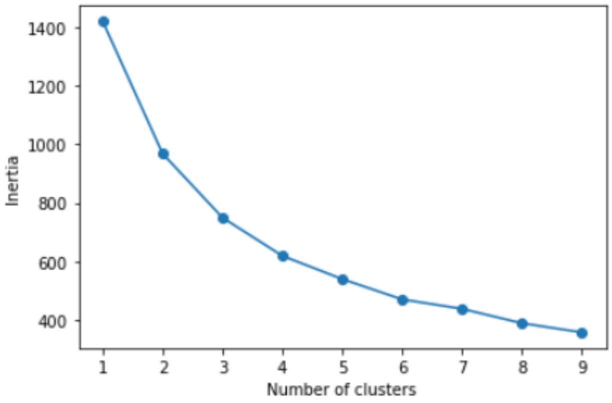
PCA reduces the dimensionality of datasets while preserving crucial information. It does this by transforming the original variables into a set of new, uncorrelated variables called principal components (vidhya, 2024).

As we analyze the variance explained by each principal component, we observe a diminishing return: additional components contribute less to the overall variance. However, after the first 5 components, the increase in cumulative explained variance becomes minimal. These 5 components collectively account for around 95% of the total variance, almost reaching 100%. There-fore 5 components will be derived. The following heat map depicts the impact of the original features on the PCA components.



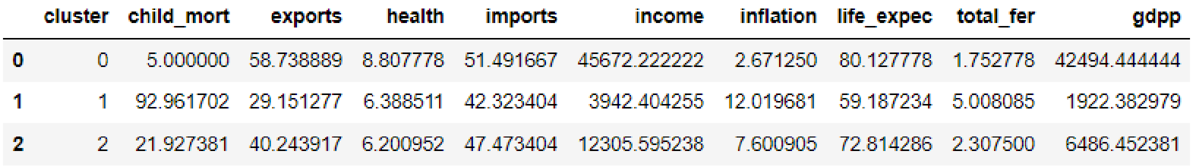


K-Means clustering

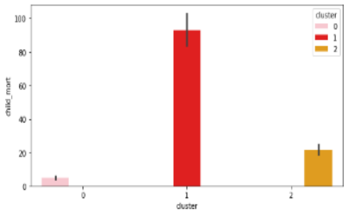


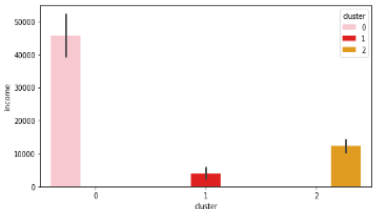
K-means clustering is a widely used method for cluster analysis where the aim is to partition a set of objects into K clusters in such a way that the sum of the squared distances between the objects and their assigned cluster mean is minimized (sharma, 2024)

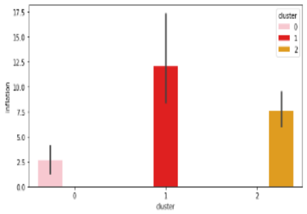
An elbow curve, which shows the within-cluster-sum-of-square (WCSS) values against its corresponding K values is plotted. The optimal K value is the point at which the graph forms an elbow (Saji, 2024). As indicated in the plot, the optimal number of clusters is 3.



	cluster	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	0	5.000000	58.738889	8.807778	51.491667	45672.222222	2.671250	80.127778	1.752778	42494.444444
1	1	92.961702	29.151277	6.388511	42.323404	3942.404255	12.019681	59.187234	5.008085	1922.382979
2	2	21.927381	40.243917	6.200952	47.473404	12305.595238	7.600905	72.814286	2.307500	6486.452381



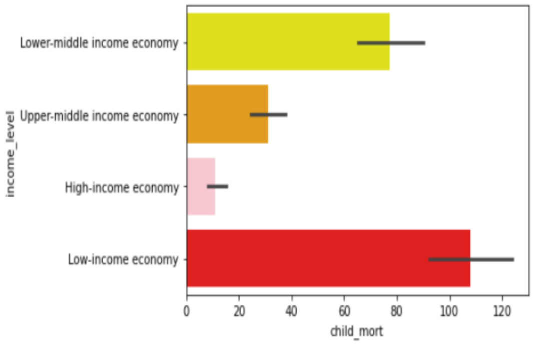




Cluster 0 - In terms of finance, countries belonging to this cluster are performing well with a favourable level of inflation, high gdp per capita and GNI indicating high standard of living and high purchasing power. In terms of health, the child mortality rate is low and spending on health care is relatively high. In terms of trade, there seems to be a surplus, hence an inflow of funds. Therefore, we can conclude that countries in this cluster don’t require aid.

Cluster 1- Countries in this cluster have the highest level of inflation on average, along with the lowest level of GDP per capita and GNI, indicating low purchasing power, standard of living. Trade seems to be unfavourable with imports being greater than exports, resulting in depletion of foreign reserves. Most importantly, child mortality is the highest, and life expectancy is the lowest suggesting that these countries are in dire need of aid. It’s also noteworthy that this cluster consists of war-tone countries such as Afghanistan, Iraq, Yemen suggesting need for humanitarian aid.

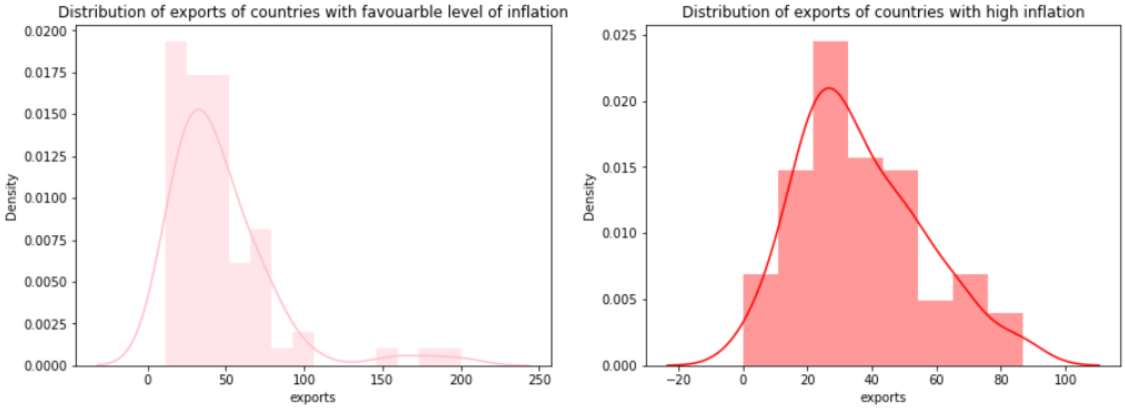
Cluster 2- In terms of finance and trade, countries in this cluster seem to be performing moderately well compared to cluster 1, however have room for improvement since inflation is still quite high. In terms of health, child mortality and life expectancy seem to be relatively well. There seems to be specific countries in this cluster that may need help, such as Argentina, Sri Lanka.



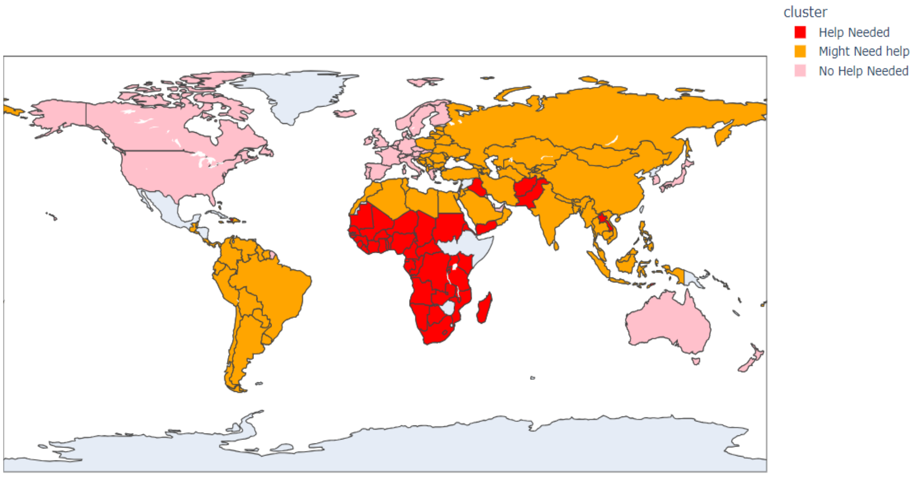
Does the income level/ GNI of a country impact the level of child mortality?

A study carried out by (O'Hare, 2013) concluded that there is an inverse relationship between income and child mortality(-0.45). In order to assess this theory, countries were grouped by their GNI into high, upper middle, lower middle and low income economies and the plot on the left was obtained. Upon examination, high income economies have the lowest level of child mortality while low-income economies have the highest. Additionally, correlation between income and child mortality is -0.52 indicating that as income increases, there is a decrease in child mortality, which aligns with the result of the study mentioned above.

Does high inflation lead to a reduction in exports?



A favourable level of inflation is considered to be 2%-3% (Ross, 2024). Therefore, countries were split based on this criterion. In the figures above, countries with high inflation show all exports being below 100, while those with favourable inflation have exports nearing or exceeding 150 highlighting a relationship between inflation and exports. Upon further examination, the correlation between the two variables is -0.11 which is a weak negative linear relationship, suggesting that as a country's inflation rises, exports fall.



Which region is currently facing the most urgent need for aid?

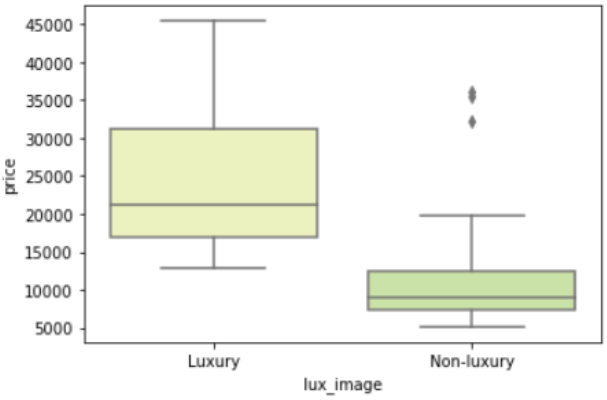
Countries from cluster 1 are predominantly located in Africa. There-fore Africa faces the most urgent need for aid. Mali, Congo, Zambia, Sudan are some countries that belong to this region. However, currently the largest recipients of foreign aid are in Sub-Saharan Africa, which happens to be where the world's lowest ranked countries in many areas of governance are, especially in terms of corruption (Lyons, 2024). Hence increased aid has resulted in increased corruption, poor wealth distribution, and aid dependence. Therefore, it's vital that aid is provided in more efficient ways that promote growth.

Task 2 - Regression

Supervised learning is a form of machine learning that uses labelled data sets to train algorithms that classify data or predict outcomes accurately (IBM). Regression is a form of supervised learning that captures the relationships between independent and dependent variables, with the main purpose of predicting an outcome (Lawton, 2023).

A dataset containing car prices (which will be considered as the dependent variable) along with categorical and continuous variables such as car brand, horsepower, and fuel type, will be utilised to answer the following research questions and identify the most accurate model to predict car price based on selected independent variables.

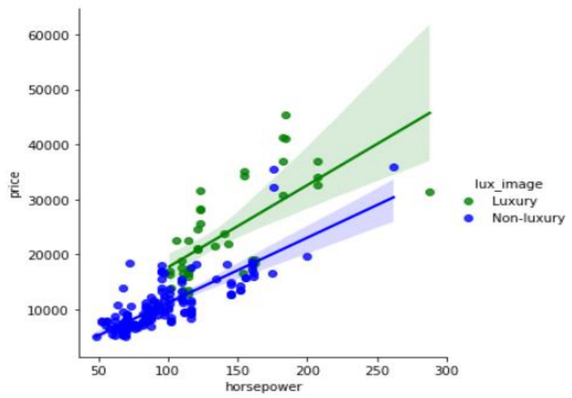
Does Cachet associated with a car brand impact the price of a car?

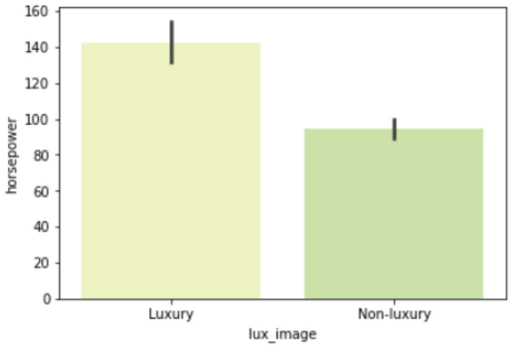


Edmunds data found that the disparity between average mainstream and luxury vehicle prices shrank about 10% between the mid-2000s and mid-2010s (Frio, 2023). In order to identify if the cachet associated with owning a luxury car still impacts price, car brands were classified into luxury and non-luxury brands and the boxplots on the right were plotted. The figure indicates that luxury vehicles have a higher spread and average price of approximately $20000, compared to the non-luxury vehicles which are priced at $10000 on average and have a maximum value of $20000, suggesting that cachet impacts price. Therefore, we can observe that companies are able to charge higher prices if their brand is perceived as luxury/ high end.

What is the relationship between horsepower of a car and its price?

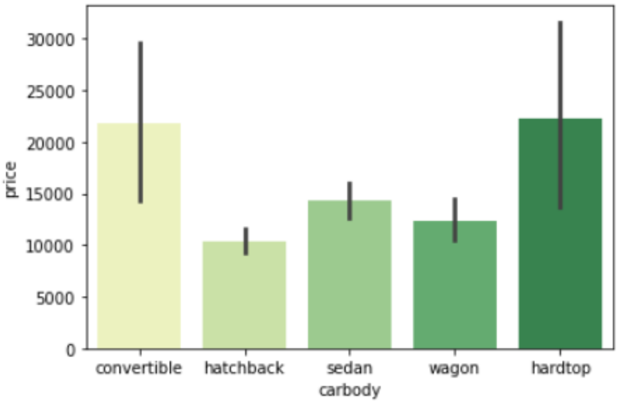
Based on the scatterplot below, horsepower and price have a positive relationship. Hence, higher the horsepower, higher the price. Correlation between the two variables is 0.81 suggesting strong association. It's noteworthy that luxury cars often exhibit higher horsepower compared to non-luxury cars, potentially serving as a confounding factor. This trend is evident in the figures depicted below.





What kind of car body is the most expensive?

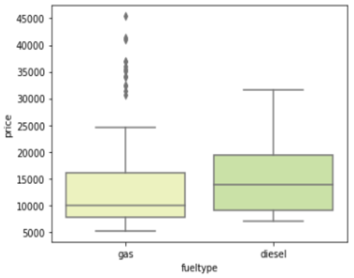
The bar plot reveals that convertibles and hardtops are the most expensive vehicles, while hatchbacks are the cheapest. This could be attributed to the higher cost than the same make and model with a standard solid top roof, since convertible roofs typically require additional support and engineering.

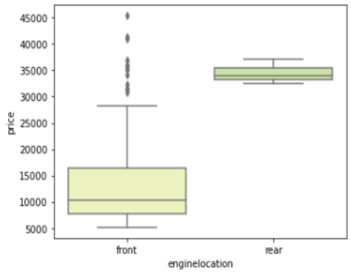


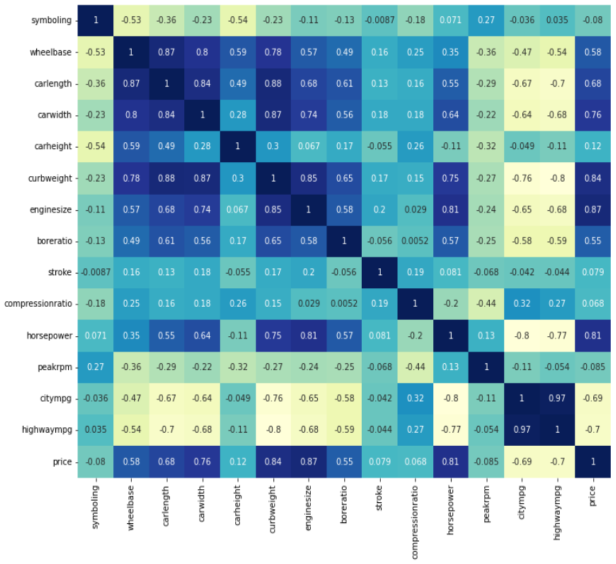
Feature selection

Categorical variables that impact the price of a car were determined by carrying out EDA, followed by assigning dummy variables to variables that showed association such as, fuel type, engine location, luxury image and features with relatively high correlation in comparison to other categorical variables were selected.
Additionally, a correlation of greater than or equal to 0.8 was assumed to be a high correlation for continuous variables. Features that had high multi-collinearity were removed such as car width.

fuel type, aspiration, engine location, cube weight, engine size, horsepower and luxury image were selected as features for predicting car prices.



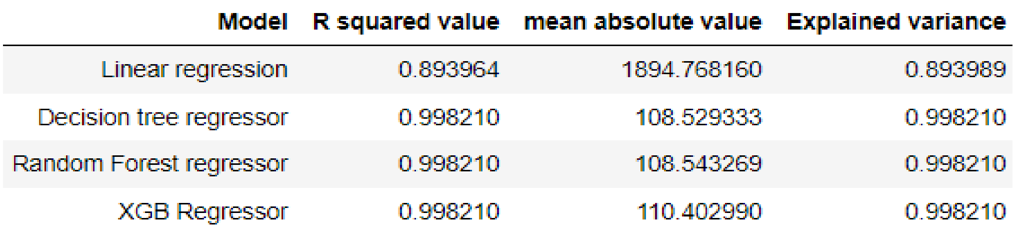




	symboling	wheelbase	carlength	carwidth	carheight	curbweight	enginesize	boreratio	stroke	compressionratio	horsepower	peakrpm	citympg	highwaympg	price
symboling	1	-0.53	-0.36	-0.23	-0.54	-0.23	-0.11	-0.13	-0.0087	-0.18	0.071	0.27	-0.036	0.035	-0.08
wheelbase	-0.53	1	0.87	0.8	0.59	0.78	0.57	0.49	0.16	0.25	0.35	-0.36	-0.47	-0.54	0.58
carlength	-0.36	0.87	1	0.84	0.49	0.88	0.68	0.61	0.13	0.16	0.55	-0.29	-0.67	-0.7	0.68
carwidth	-0.23	0.8	0.84	1	0.28	0.87	0.74	0.56	0.18	0.18	0.64	-0.22	-0.64	-0.68	0.76
carheight	-0.54	0.59	0.49	0.28	1	0.3	0.067	0.17	-0.055	0.26	-0.11	-0.32	-0.049	-0.11	0.12
curbweight	-0.23	0.78	0.88	0.87	0.3	1	0.85	0.65	0.17	0.15	0.75	-0.27	-0.76	-0.8	0.84
enginesize	-0.11	0.57	0.68	0.74	0.067	0.85	1	0.58	0.2	0.029	0.81	-0.24	-0.65	-0.68	0.87
boreratio	-0.13	0.49	0.61	0.56	0.17	0.65	0.58	1	-0.056	0.0052	0.57	-0.25	-0.58	-0.59	0.55
stroke	-0.0087	0.16	0.13	0.18	-0.055	0.17	0.2	-0.056	1	0.19	0.081	-0.068	-0.042	-0.044	0.079
compressionratio	-0.18	0.25	0.16	0.18	0.26	0.15	0.029	0.0052	0.19	1	-0.2	-0.44	0.32	0.27	0.068
horsepower	0.071	0.35	0.55	0.64	-0.11	0.75	0.81	0.57	0.081	-0.2	1	0.13	-0.8	-0.77	0.81
peakrpm	0.27	-0.36	-0.29	-0.22	-0.32	-0.27	-0.24	-0.25	-0.068	-0.44	0.13	1	-0.11	-0.054	-0.085
citympg	-0.036	-0.47	-0.67	-0.64	-0.049	-0.76	-0.65	-0.58	-0.042	0.32	-0.8	-0.11	1	0.97	-0.69
highwaympg	0.035	-0.54	-0.7	-0.68	-0.11	-0.8	-0.68	-0.59	-0.044	0.27	-0.77	-0.054	0.97	1	-0.7
price	-0.08	0.58	0.68	0.76	0.12	0.84	0.87	0.55	0.079	0.068	0.81	-0.085	-0.69	-0.7	1

Regression models

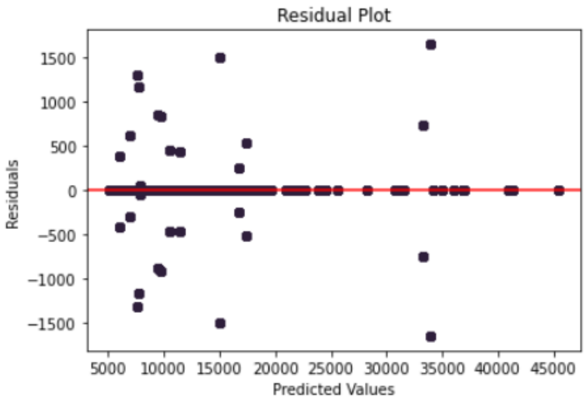
The dataset was fitted using multiple models in order to identify the most accurate model to predict car prices. Data was scaled according to standard scaler, where features were standardized, and split into training and testing sets by a 80:20 ratio since the dataset is quite small. Multiple linear regression, Decision tree regressor, random forest regressor and XGBoost regressor were considered and the following results were obtained.



Model	R squared value	mean absolute value	Explained variance
Linear regression	0.893964	1894.768160	0.893989
Decision tree regressor	0.998210	108.529333	0.998210
Random Forest regressor	0.998210	108.543269	0.998210
XGB Regressor	0.998210	110.402990	0.998210

R-squared value - R-squared is a statistical measure that indicates how much of the variation of a dependent variable is explained by an independent variable in a regression model (Fernando, 2023).
Mean absolute error-shows the average distance of the observation of the dataset from the mean of the dataset (Academy).
Explained variance- This measures the variance between a model and the actual data (Capital, 2024).

Multiple linear regression model has the lowest R^2 value and therefore is the least accurate model. It’s noteworthy that the latter three models are almost identical in their performance, with minimal deviation in their respective evaluation scores. Since Decision tree regressor has the lowest MAE, it will be considered as the best fit for predicting car prices. 99.8% of the variation in car prices is due to the variation in its features, which is favourable. Hyper parameter tuning was carried out to improve the accuracy further, however the original model was deemed as more accurate. The residual plot depicting the error terms is indicated above. We can observe that the error terms are homoscedastic and there is no autocorrelation.

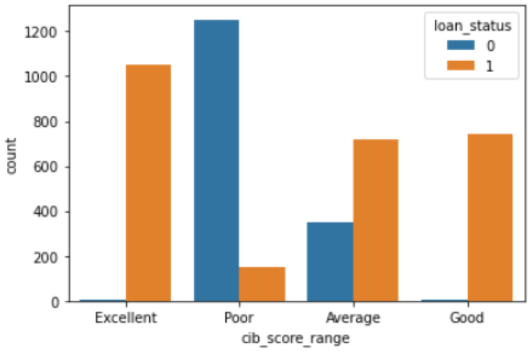


Task 3 - Classification

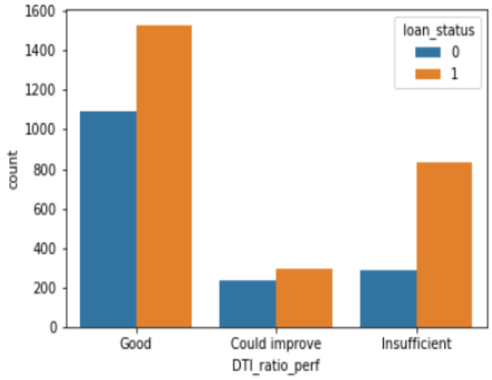
Classification is another form of supervised machine learning that involves predicting the class of given data points. Those classes can be targets, labels or categories (IBM).

The dataset used for this task contains financial records and information used to determine the eligibility of individuals or organizations for obtaining loans from a lending institution. This falls under binary classification as the output variable that needs to be predicted is discrete and falls into 2 classes: loan approve (1) or not approved (0).

Does one's CIBIL score impact the likelihood of loan approval?



79% of the loans that are approved are to the individuals who have a Credit score that is greater than 750 (bazaar). In order to examine this finding, data was grouped based on their respective cibil score ranges and the bar graph was derived. We can observe how individuals with good (650-750) and excellent (750-900) cibil scores get their loans approved majority of the time whereas poor cibil scores result in predominantly rejected loans. This indicates that cibil score impacts the likelihood of getting a loan approved, which aligns with the findings above. Additionally, cibil score and loan approval have a correlation of 0.77 which suggests strong positive linear relationship.

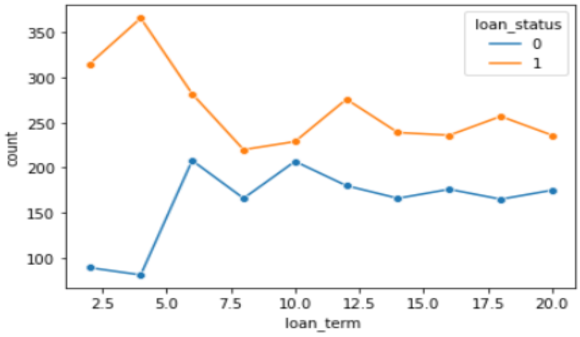


Does a low DTI ratio affect an individual's eligibility of being approved for a loan?

DTI ratio is a personal finance measure that compares the amount of debt you have to your overall income (consumer financial protection bureau, 2023).

DTI ratio = Gross monthly debt / Gross monthly income

Loan amount per month was considered as gross monthly debt and DTI ratio was calculated. Based on the figure, individuals with a good (<35%) DTI ratio have the largest number of loan approvals. However, they also have a large number of loans being rejected. This suggests that having a favourable DTI score doesn’t guarantee the loan being approved. Furthermore, DTI ratio and loan status have a correlation of 0.18 which is a weak positive linear relationship.



What duration of loan terms has the highest likelihood of getting approved?

The line chart indicates that 2-4 year loan terms have the highest possibility of being approved. This could be attributed to financial institutions justifying charging higher interest rates on short-term loans compared to long-term loans by citing the increased risk associated with shorter repayment periods. Consequently, they are more inclined to approve short-term loans, as they offer a greater potential for profit through higher interest charges.

Feature selection

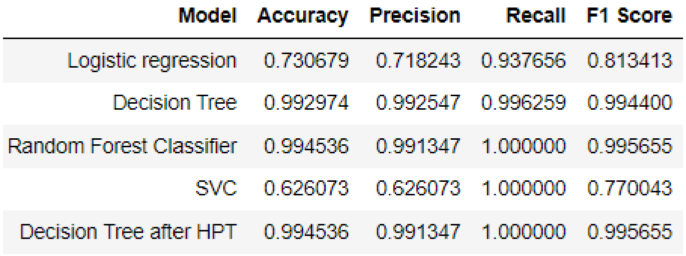
Besides Cibil score, correlation of other variables with the dependent variable are quite weak. Therefore, features were primarily selected based on variables that indicated association with loan status in EDA and variables with correlation greater than 0.01.

Number of dependents, income, loan amount, loan term, cibil score, residential assets value, DTI ratio were selected as independent variables to predict loan status.

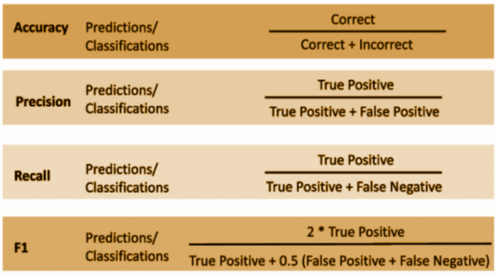
Classification Models

The dataset was split into a training and testing set based on a 70:30 ratio and fitted on the following models in order to identify which model is the most accurate to predict loan status.

Logistic regression - the model estimates the probability of an event occurring, based on a given data set of independent variables (IBM).
Decision Tree classifier- It's an algorithm that uses a set of rules to make decisions, similarly to how humans make decisions.
Random Forest classifier- This model combines the output of multiple decision trees to reach a single result.
Support vector classification- The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N -the number of features) that distinctly classifies the data points.



Model	Accuracy	Precision	Recall	F1 Score
Logistic regression	0.730679	0.718243	0.937656	0.813413
Decision Tree	0.992974	0.992547	0.996259	0.994400
Random Forest Classifier	0.994536	0.991347	1.000000	0.995655
SVC	0.626073	0.626073	1.000000	0.770043
Decision Tree after HPT	0.994536	0.991347	1.000000	0.995655

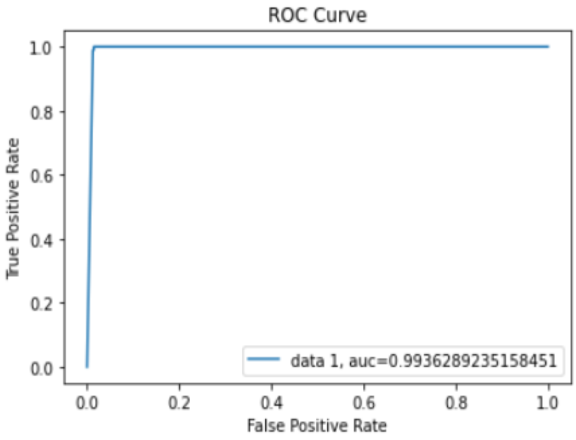


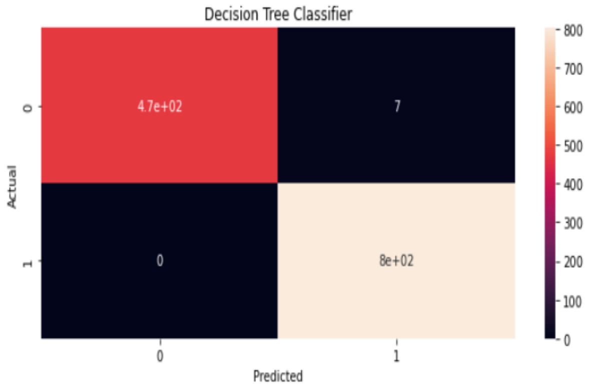
Accuracy	Predictions/ Classifications	$\frac{\text{Correct}}{\text{Correct + Incorrect}}$
Precision	Predictions/ Classifications	$\frac{\text{True Positive}}{\text{True Positive + False Positive}}$
Recall	Predictions/ Classifications	$\frac{\text{True Positive}}{\text{True Positive + False Negative}}$
F1	Predictions/ Classifications	$\frac{\text{2 * True Positive}}{\text{True Positive + 0.5 (False Positive + False Negative)}}$

62.2% of the data consists of loans that are approved whereas 37.8% are not approved. Since the data is imbalanced, accuracy of the models maybe misleading therefore Recall and F1 score will be used to determine the best model.

Logistic regression and SVC relatively have low precision and relatively low F1 score. Therefore Decision tree and random forest classifier are considered to be the best models. In order to

Optimize the models further, hyper-parameter tuning was carried out. Decision tree model saw an improvement where recall increased to 1 and the accuracy scores were identical to random forest classifiers while random forest model remained the same.





We can observe how the ROC curve of the Decision tree model is closest to the top left corner with an AOC of 0.9934 and there-fore is the best model. Additionally, out of the total loans rejected, only 7 were inaccurately predicted as approved, while none of the loans that were actually approved were mistakenly predicted as rejected.

References

Academy, K. (n.d.). Retrieved from https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/other-measures-of-spread/a/mean-absolute-deviation-mad-review

bazaar, b. (n.d.). Retrieved from https://www.bankbazaar.com/cibil/loan-approval-process-on-cibil-score.html

Capital, F. (2024). Explained variance: The Role of R squared in Quantifying Model Performance. Retrieved from https://fastercapital.com/content/Explained-variance--The-Role-of-R-squared-in-Quantifying-Model-Performance.html#:~:text=Explained%20variance%20is%20a%20measure,mean%20of%20the%20dependent%20variable.

consumer financial protection bureau. (2023). What is a debt-to-income ratio? Retrieved from https://www.consumerfinance.gov/ask-cfpb/what-is-a-debt-to-income-ratio-en-1791/#:~:text=Your%20debt%2Dto%2Dincome%20ratio,will%20have%20different%20DTI%20limits.

Fernando, J. (2023). R-Squared: Definition, Calculation Formula, Uses, and Limitations. Retrieved from https://www.investopedia.com/terms/r/r-squared.asp

Frio, D. (2023). Top Luxury Car Brands of 2023. Retrieved from https://www.edmunds.com/car-buying/luxury-car-brands.html

IBM. (n.d.). What is ML? Retrieved from https://www.ibm.com/topics/machine-learning#:~:text=Unsupervised%20learning%2C%20also%20known%20as,the%20need%20for%20human%20intervention.

IBM. (n.d.). What is supervised learning? Retrieved from https://www.ibm.com/topics/supervised-learning#:~:text=Supervised%20learning%20uses%20a%20training,error%20has%20been%20sufficiently%20minimized.

Lawton, G. (2023). What is regression in machine learning? Retrieved from https://www.techtarget.com/searchenterpriseai/feature/What-is-regression-in-machine-learning#:~:text=Regression%20in%20machine%20learning%20is,distribution%20of%20each%20data%20point.

Lyons, J. (2024). Foreign aid is hurting, not helping Sub-Saharan Africa. Retrieved from https://www.lejournalinternational.fr/Foreign-aid-is-hurting-not-helping-Sub-Saharan-Africa_a2085.html

O'Hare, B. (2013). *Income and child mortality in developing countries: a systematic review and meta-analysis.* Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3791093/

Ross, S. (2024). When Is Inflation Good for the Economy? Retrieved from https://www.investopedia.com/ask/answers/111414/how-can-inflation-be-good-economy.asp

Saji, B. (2024). Elbow Method for Finding the Optimal Number of Clusters in K-Means. Retrieved from https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/

sharma, p. (2024). The Ultimate Guide to K-Means Clustering: Definition, Methods and Applications. Retrieved from https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

vidhya, a. (2024). PCA | What Is Principal Component Analysis & How It Works? (Updated 2024). Retrieved from https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/