ST3189

MACHINE LEARNING

Unsupervised learning, Regression and Classification

Table of Contents

This report should be read in conjunction with the python code file in the Jupyter notebook named 200643540_ST3189.ipynb as the report demonstrates the results from the tasks carried out in python.

Unsupervised Learning

Unsupervised learning refers to the machine learning techniques that deal with unlabelled datasets, that is datasets with no prespecified target column. Clustering is an unsupervised machine learning model which uses the similarities in features on offer to group the data points and generate an insightful output. In this task, the K-Means clustering algorithm is used to analyse bank_authentication_notes.csv available at Kaggle and UCI Machine Learning Repository: banknote authentication Data Set.

K-Means

K-Means is an algorithm used to partition the dataset into k number of prespecified clusters as per requirement, such that intracluster similarity is high but the intercluster similarity is low. Each data point will belong to one, and only one cluster and that cluster will be the one which is the closest in distance to the cluster centroids.

Banknote authentication

Counterfeit banknotes are currency produced illegally without the proper authority by the state in an attempt to imitate legal currency and deceive recipients. They pose a significant threat to states, banks, businesses, and the general public. Maintaining confidence in the currency is one of the key obligations of the authorities, and over the years many advanced security features have been built into notes to safeguard against counterfeits (CBSL, 2018). However, counterfeiters have found ways, particularly with the advancement in printing technology (Shahani et al., 2018), to bypass those security measures in place. This means that detection of counterfeits has become an integral part in the battle against counterfeit banknotes. These detection tools have been rolled out at crucial points in currency circulation such as bank teller counters and lately cash deposit machines (Ragavi, 2020).

Higher denomination currency notes are prone to higher risk of counterfeiting (Shahani et al., 2018) as well as higher damages to the society when done. Counterfeiting often traces back to money laundering and underground economic activities which pose a threat to society in many other ways. Increased awareness and shift towards digital transactions have curbed the risk of counterfeits but it still remains a significant issue in need of urgent addressing.

The data in the dataset were extracted from images that were taken from genuine and forged banknotes. Regions of interest (ROI) – (Figure1), were selected accounting for homogeneity of the textures. The Wavelet Transform tool was used to extract features from images (Lohweg & Gillich, 2010).

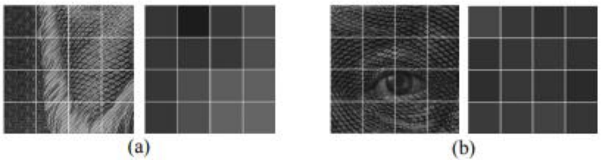


Figure 1 - (Lohweg & Gillich, 2010)

The following are a non-exhaustive list of research questions pertaining to the study of counterfeit notes for which this project and further research can aim to find solutions.

RQ1. Are there any relationships between the extracted features?
RQ2. Do the features extracted show any distinct patterns to enable grouping?
RQ3. Can images of banknotes be used for authentication purposes?
RQ4. Do built in advanced security features make it difficult to produce counterfeits?
RQ5. Does higher denomination currency counterfeiting link to money laundering and other underground economic activities?

Exploratory Data Analysis (EDA)

The dataset consisted of 5 attributes, namely, Variance of Wavelet Transformed image (continuous), Skewness of Wavelet Transformed image (continuous), Curtosis of Wavelet Transformed image (continuous), Entropy of image (continuous), and Class (integer). Since the first model makes use of unsupervised learning, the column named 'Class', which consisted of the labels was dropped from the dataset.

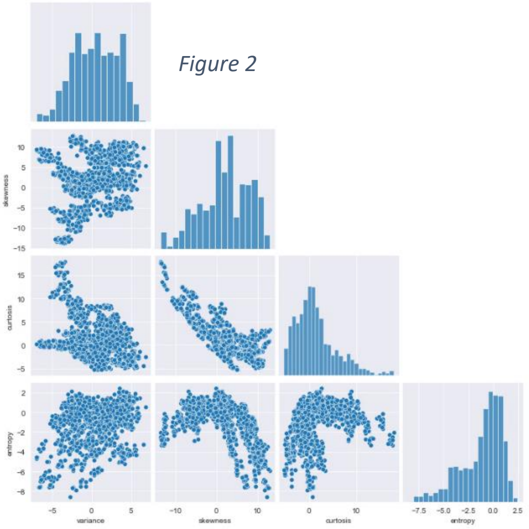


Figure 2

Through EDA several observations were made on the distribution of the features and their relationships with each other. The pairwise plot (Figure2) visualises both the distribution of each variable and their respective relationships with other variables in the dataset. In this instance, we can see an almost linear relationship with a negative slope between skewness and kurtosis. Entropy seems to have a certain hyperbolic relation with skewness and kurtosis.

Variance and skewness are evenly distributed while kurtosis and entropy seem to have values that lie on extreme ends giving them positively and negatively skewed distributions respectively. This is visualised through the boxplots, violin plots as well as the outer diagonal of the pairwise plot.

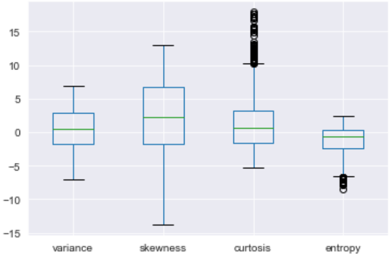


Figure 3

From the correlation heat map (Figure 5), we can see that skewness and kurtosis are highly correlated along with entropy and skewness which are moderately correlated.

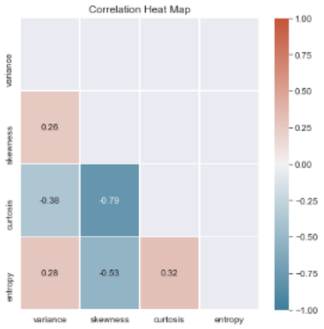


Figure 5

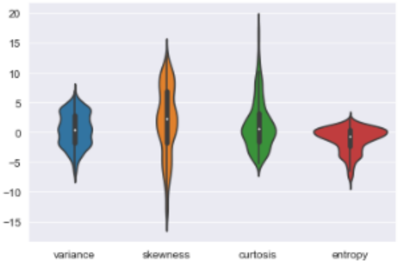
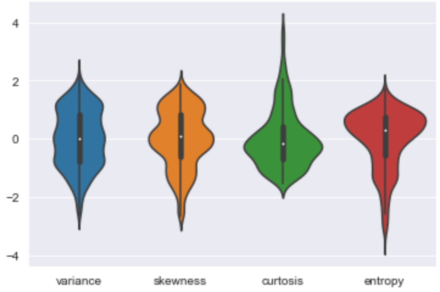


Figure 4

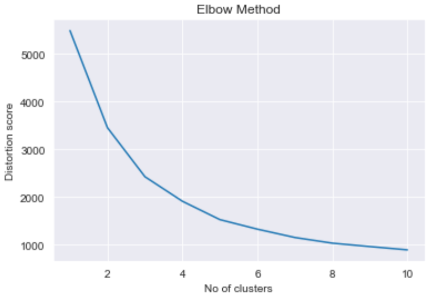
Model

The process of training this model and all others starts with importing the necessary libraries which allow the relevant datasets to be transformed for analysis. For this question, NumPy, pandas, matplotlib, seaborn and sklearn were imported. These libraries contain chunks of reusable code that aid in data manipulation, data visualization, and the building of the model.

Figure 6



The dataset was checked for null values and there were none. Upon describing all the features of the dataset, the numeric features were standardised so that all features influence the model similarly. If not, the model gets influenced more by features with a larger scale and behave badly. A much smoother violin plot was observed post standardisations compared to earlier (Figures 4 & 6).



The Elbow method, which finds the 'Within-Cluster Sum of Squares', was used to determine the optimal number of clusters (value of k) into which the data may be clustered. At the value of k=2, a linear decrease of distortion was observed.

Figure 7 - Elbow method

Then the K-Means model was fitted, using k=2, clustering the data into two distinct groups (Figure 8). These clusters were visualised and a clear differentiation was observed. Experimenting with k=3, 4 – a higher level of overlap in clusters was observed thus proving that k=2 was the optimal number of clusters (Figures 9 & 10).

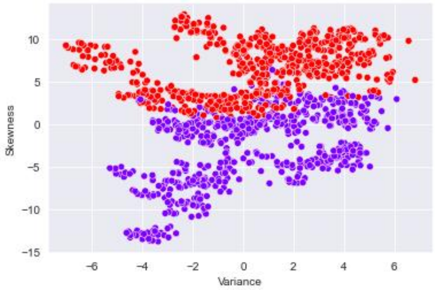


Figure 8 – (K = 2)

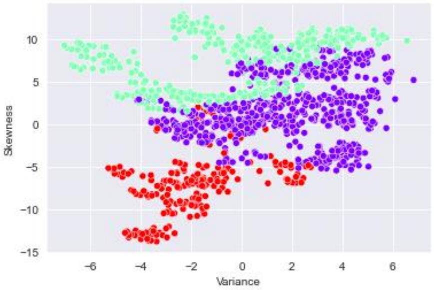


Figure 9 – (K = 3)

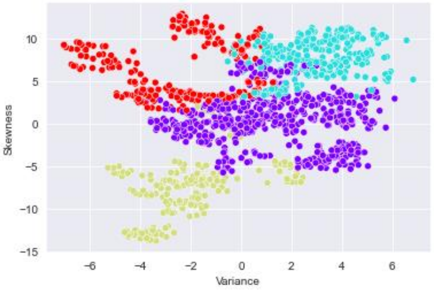


Figure 10 – (K = 4)

Principal component analysis, or PCA is another unsupervised model used as a dimensionality reduction technique. It is used to reduce multiple columns of the dataset into a few without losing the information containing in them. In this case, all the four attributes of the dataset were reduced to two, since it was determined that >80% of the variance was explained by 2 PCs.

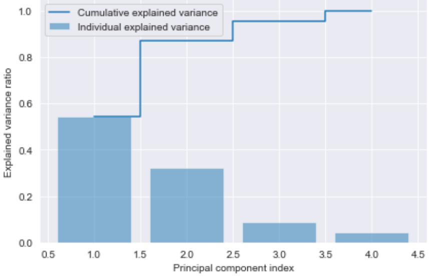


Figure 11 – Explained variance ratio PCA

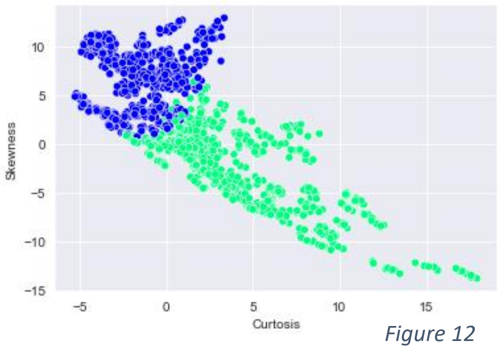


Figure 12

The K-Means model was run again to compare it against the first attempt. With PCA, the clusters showed a similar distinction, proving that the two principal components were in fact capable of recognising the patterns for partitioning (Figure 12).

How the four features were distributed among the clusters was evaluated using boxplots separated by the cluster assigned. There were some notable relationships visualised, for instance, cluster = 0 favoured records with higher skewness and cluster = 1 favoured records with lower skewness (Figure 13). The contrary was true for kurtosis and entropy.

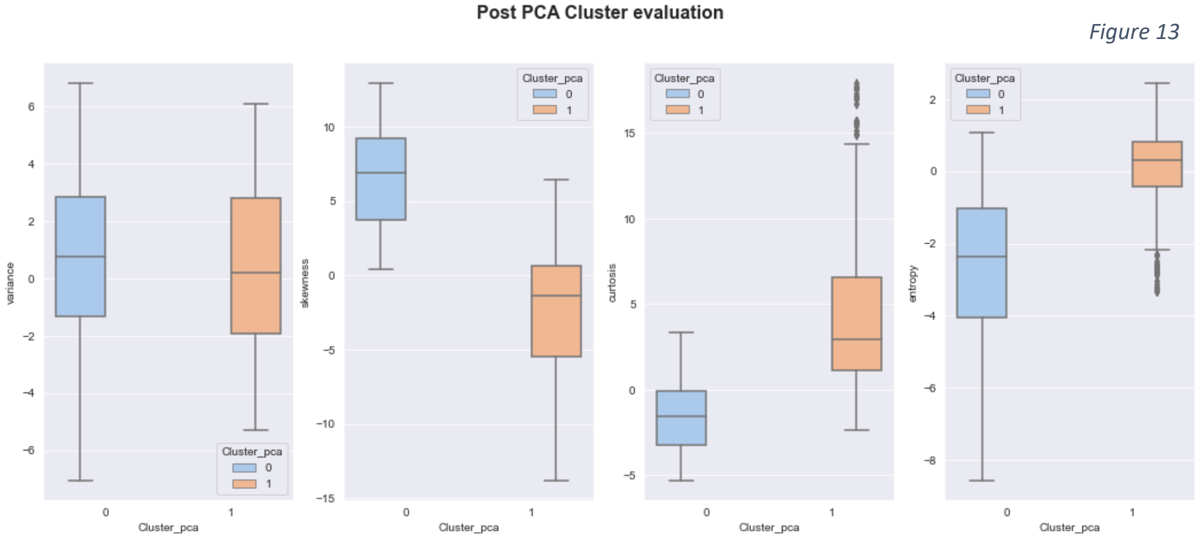


Figure 13

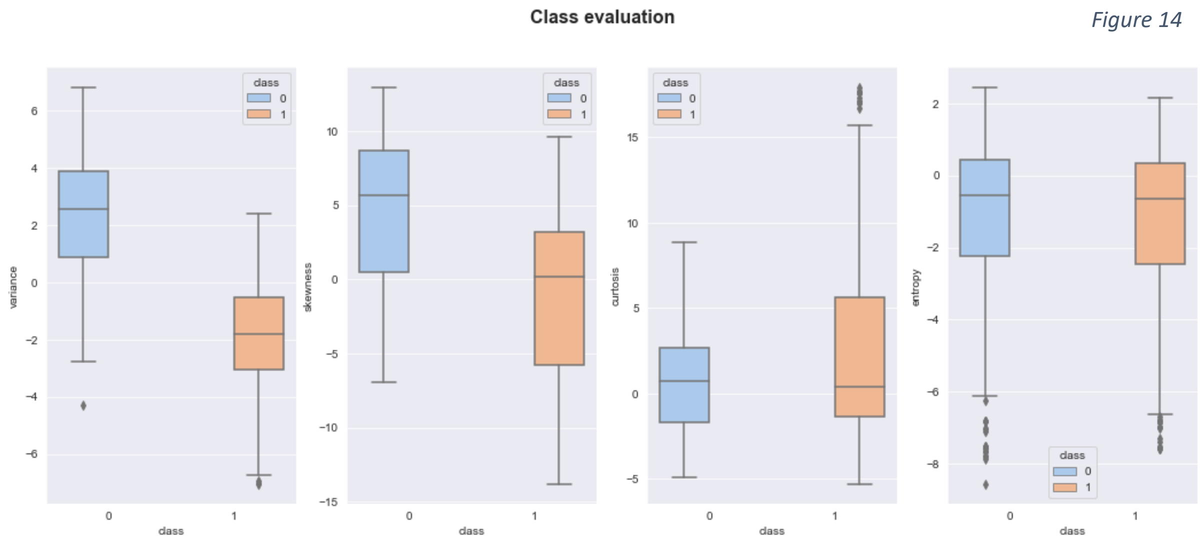
Classification

Classification is a supervised machine learning technique that deals with datasets with labelled, categorical target (dependent variable) columns. Classification models learn patterns from the labelled input data, also known as the training set and uses that learning to predict the classes of new observations.

Exploratory Data Analysis (EDA)

For this task, we will be using the same dataset used for the unsupervised model but this time with the labels. The data pre-processing would take a process similar to the first model since it is the same dataset. Since no exploratory analysis was done on the classes before, it was evaluated before running the classification models. The dataset contains a balanced ratio of the two classes at 56:44 (genuine = 0 : counterfeit = 1). Features show a distinct distribution in each class, as depicted by the boxplots (Figure 14).

Figure 14



Models

Four different classification models, namely, Gaussian Naive Bayes, Decision Tree, Logistic Regression, and Neural Network were trained parallelly to compare each of their performances. All of the models were imported from different verticals of the sklearn library.

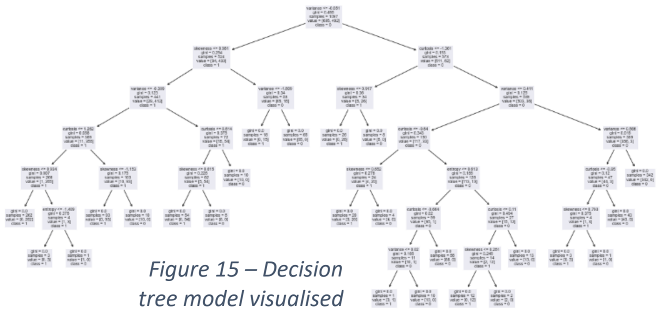


Figure 15 – Decision tree model visualised

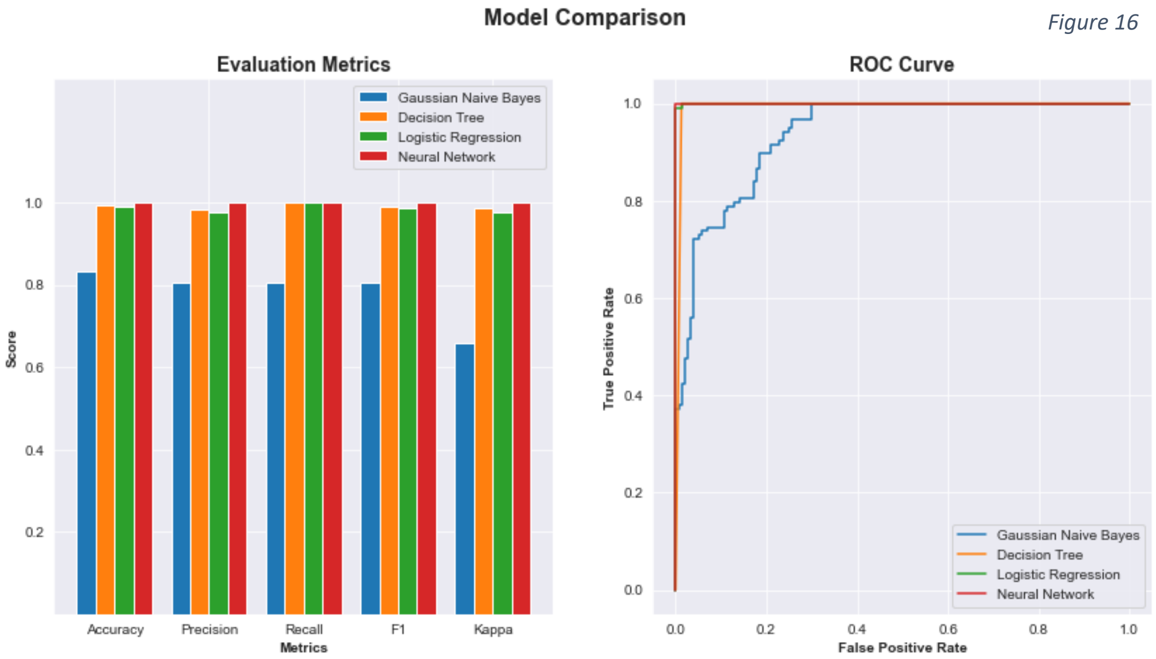
The dataset was first split into features – independent variables and labels – dependent variables. Then it was further split into training and testing, 80% and 20% respectively. The training set is the input of the model, from which it learns the patterns and maps a function of independent variables to the discrete output variable - the independent, labelled, target. The test set is data held back from the training of the models, so that it can be used to estimate the performances of the models post training.

The standardising was done this time too, but unlike before it was done separately for the test set and training set to avoid any data leakage. Data leakage is when knowledge of the test set leaks into the dataset used to train the model which can result in an incorrect estimate of model performance.

A function was defined to ease the process of evaluating each of the models. Performance was measured in terms of accuracy, precision, recall, f1-score, and kappa score. The confusion matrix and the AUC (area under curve) was also produced to visually project the performances.

Gaussian Naive Bayes was used as a baseline model, and as expected, it performed the worst out of the four, but still at a reasonable precision level of 80%. Decision Tree, Logistic Regression, and Neural Network all three performed impressively well, clocking in accuracy and precision at almost 100%. In fact, the Neural Network produced a perfect model which will be unseen in models trained with larger, unclean, and complex datasets (Figure 16).

Figure 16



Since the performance of the models were up to mark, no further tuning of the hyperparameters were deemed necessary.

Conclusions

RQ1 – Through EDA the existing relationships among the four features were discussed.

RQ2 – The distinct clusters generated through K-Means and the subsequent cluster evaluation established that the features showed distinct patterns capable of grouping.

RQ3 – It was concluded that the images of banknotes can in fact be used for authentication purposes given the high accuracy and precision recorded in the classification task.

Setting up these models along with devices capable of extracting the same features in the dataset from the note under observation can accurately classify between forged and real banknotes, thus helping in the battle against counterfeit notes.

The data is insufficient to conclude on RQ4 and RQ5 as the dataset doesn't contain any data pertaining to the denomination of the currency, the level of inbuilt security features or their origin. Further research and larger sampling of the population may help answer these questions in the future.

Regression

Regression is a supervised machine learning technique that deals with datasets with continuous target (dependent variable) columns. It models the relationship between a dependent variable (target) and independent (predictor) variables, with the objective of predicting a continuous outcome for changes in the independent variables. Some predictors have a linear relationship with the target while others have a nonlinear relationship. In this task, Linear Regression and Gradient Boosting are used to build regression models for the Metro_Interstate_Traffic_Volume.csv available at UCI Machine Learning Repository: Metro Interstate Traffic Volume Data Set.

Metro Interstate Traffic Volume

Traffic is an issue that has plagued the modern civilisation for a long time. It wastes valuable resources like time and energy and contributes to environmental pollution. With the increase of population and subsequent increase in vehicles on the road, traffic has only worsened in the past few years. States and cities continue to try innovative methods to curb the nuisance caused by traffic to commuters, however most have not delivered the expected results.

The data in the dataset are of two kinds; Traffic data and Weather data. Traffic data was extracted from the MN Department of Transportation and Weather data from the OpenWeatherMap. It contains hourly Interstate 94 Westbound traffic volume for MN DoT ATR station 301, roughly midway between Minneapolis and St Paul, MN and hourly weather features. Holiday status of the day is also included (Hogue, 2019).

The following are a non-exhaustive list of research questions pertaining to the study of counterfeit notes for which this project and further research can aim to find solutions.

RQ1 – Has traffic volume increased over the years?

RQ2 – Is there noticeable seasonality in the traffic volume?

RQ3 – Can weather patterns and holiday status accurately predict traffic volume?

Linear Regression

Linear regression is a supervised learning model which attempts to predict a target value based on a linear function of the independent variables. In other words, it tries to calculate the target with a combination of the predictor variables and their respective weights which we call coefficients. Intercept is the part of the model that is not affected by the independent variables and is a constant.

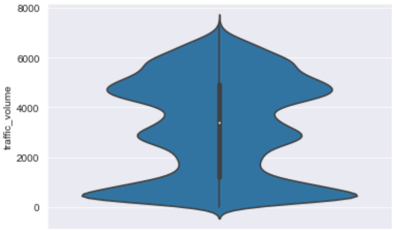
Gradient Boosting

Gradient boosting builds an additive model in a forward stage-wise fashion which allows for the optimization of arbitrary differentiable loss functions. A regression tree is fit on the negative gradient of the given loss function at each stage (Sklearn, 2023). It is a powerful model capable of finding non linear relationships between the target and the given features while dealing well with high cardinality categorical values on features.

Exploratory Data Analysis (EDA)

The dataset consisted of 9 attributes, namely, ‘holiday’, ‘temp’, ‘rain_1h’, ‘snow_1h’, ‘clouds_all’, ‘weather_main’, ‘weather_description’, ‘date_time DateTime’, and ‘traffic_volume’. Each of them describes conditions under which the target – traffic_volume was observed, such as weather conditions and timing. A full description of the dictionary is available at the source of the dataset.

Figure 17



Distribution of the target – traffic volume was visualised using a violin plot to get an understanding of the quartiles and the density of the data (Figure 17).

Then the traffic volume data was grouped yearly, monthly, hourly and by holiday status. The groups were averaged and then plotted simultaneously with the mean traffic volume of the entire target column (Figures18,19 & 20). It was observed that the volume of traffic had a downward trend from 2013-2016, but increased to above average levels in 2017 and afterwards. The deviations are small in volume and close to the average hence it was concluded that a stable level of traffic was maintained over the years.

However, monthly and hourly traffic data shows strong signs of seasonality. December has recorded the lowest volumes of traffic owing possibly to the holiday season where commute to work and school is minimal. Hourly data shows that highest level of congestion occurs during 1500-1600, which is when most commuters get off work and start their journey back home. Traffic volume is also high during the morning rush hour from 0700. From 2300-0400, the roads are not filled with traffic as those are the inactive times for most.

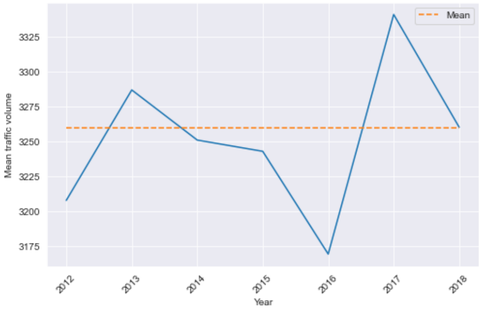


Figure 18 - Yearly

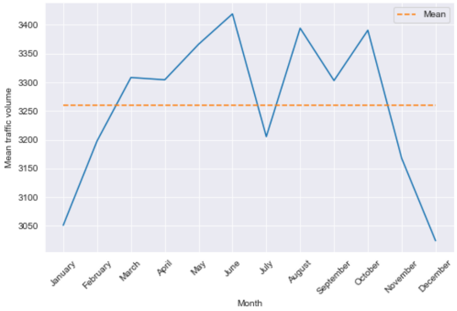


Figure 19 - Monthly

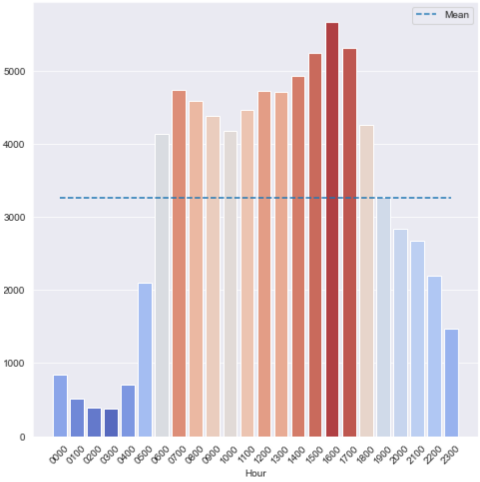
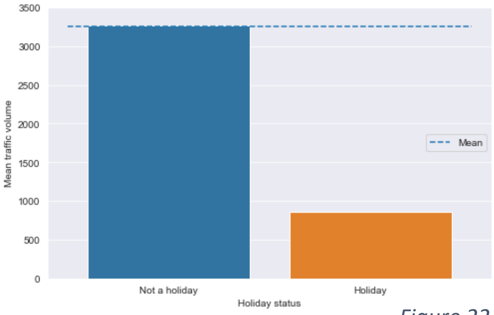


Figure 20 - Hourly

Figure 21 – Traffic volume by holiday status



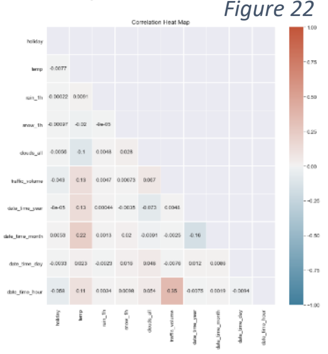
Holiday status has a strong relationship with traffic, where the volume of traffic is drastically low during holidays and up to average levels on other days, thus confirming our previous hypothesis on work and school commute contributing vastly towards road congestion.

Models

Apart from the libraries imported for data manipulation and data visualization in the prior models, One Hot Encoder was imported.

The dataset was checked for null values and there were none. Correlation amongst features was also minimal (Figure 22).

Figure 22



Categorical variables have to be specially coded to be used in regression, particularly nominal categorical variables. This is because most machine learning models accept and understand only numerical variables.

In this data set, three of the columns had to be encoded prior to training the model. Through a value count of the unique values in the ‘holiday’ column, it was noted that it had mostly ‘None’ values and a few of the holidays that came around only once a year. Hence, it was transformed to a two-class feature, 0 – No holiday and 1 – Holiday. This was done through a basic function to assign 0 to rows that corresponded to ‘None’ and 1 to all else.

For the other two categorical variables - 'weather_main' and 'weather_description', one hot encoding was used. One-Hot Encoding negates any effects of the model assuming an ordinal relationship, which may happen if integer encoding is used. It gives binary values, also known as dummy values to each of the different categories available as they transform into their own columns. This process largely increases the dimensions of the dataset and can bring about the curse of dimensionality. The curse of dimensionality refers to the difficulties an algorithm faces when working with datasets of the higher dimensions compared to lower dimensions. This is because higher dimensions mean higher number of combinations, which significantly raises the minimum data required to be seen by the model in order to perform well. However, in certain cases like this one, it cannot be avoided.

The dataset was then split into features and labels, and further into training and testing just like in the previous models. Numerical features were standardised to avoid distortion in the model. And a function was written to ease the process of evaluating models.

The first model trained was a linear regression model, which performed extremely poorly proving that the relationship between the features and the target is in fact nonlinear.

GradientBoostingRegressor was then trained as the second model, which performed reasonably well as expected, given that it detects nonlinear relationships.

GradientBoostingRegressor underwent hyperparameter tuning to increase its performance. Through grid search and multiple fold cross validation, the optimal combination of hyperparameters was obtained. The tuned model performed marginally better than the original one and the full comparison of the R-squared score and error levels were then plotted (Figure 23). R-Squared (R^2) is an evaluation measure in a regression model that determines the proportion of variance in the target that is explained by the model (higher the better). The various measures of error calculates the difference between the actual value and the predicted value, and these are measures that models try to minimise as much as possible.

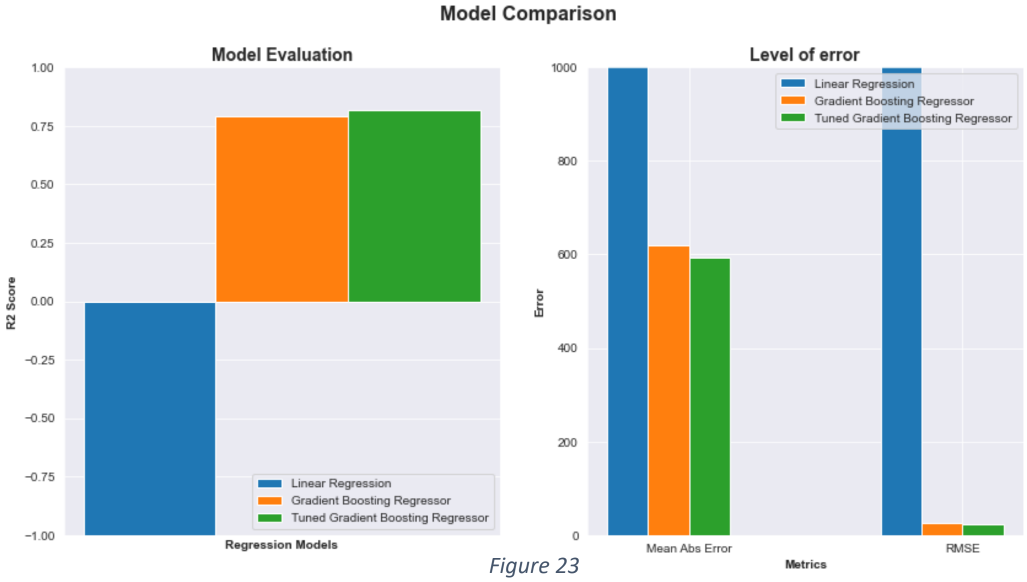


Figure 23

Conclusions

RQ1 – There is no substantial evidence to conclude that traffic has increased over the years, it has remained relatively stable.

RQ2 – There is strong evidence to suggest seasonality in traffic volume as explained in EDA.

RQ3 – Weather patterns and holiday data can reasonably predict traffic volume with a small error. The best model was able to explain the variance of traffic volume up to 81% and an RMS error of 24 – (GradientBoostingRegressor post tuning)

Predicting traffic volume ahead of time can help cities disburse their resources effectively to manage it. Which can ultimately help people to suffer less while helping the environment with reduced emissions. Further research must be done on new developments in commute such as the increased remote working culture, the use of self-driving electric vehicles, and ride sharing platforms to gain more timely insights on road congestion.

References

CBSL (2018) *Counterfeit prevention: Central bank of sri lanka*, *Counterfeit Prevention | Central Bank of Sri Lanka*. Available at: https://www.cbsl.gov.lk/en/notes-coins/damaged-notes-and-counterfeits/counterfeit-prevention (Accessed: March 25, 2023).

Chandel, A. and Sagar, S. (2020) “An accurate estimation of interstate traffic of Metro City using linear regression model of machine learning,” *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.3598310.

Hogue, J. (2019) *Metro Interstate Traffic Volume Data Set*, *UCI Machine Learning Repository: Metro Interstate Traffic Volume Data Set*. Available at: https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume (Accessed: April 2, 2023).

Hogue, J. (2019) *Metro Interstate Traffic Volume Data Set*, *UCI Machine Learning Repository: Metro Interstate Traffic Volume Data Set*. Available at: https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume (Accessed: March 27, 2023).

Hunter, J.D. (2007) “Matplotlib: A 2D graphics environment,” *Computing in Science & Engineering*, 9(3), pp. 90–95. Available at: https://doi.org/10.1109/mcse.2007.55.

Lohweg, V. and DÃrksen, H. (2013) *UCI Machine Learning Repository: Banknote Authentication Data Set*. Available at: https://archive.ics.uci.edu/ml/datasets/banknote+authentication (Accessed: March 25, 2023).

Lohweg, V. and Gillich, E. (2010) “Banknote Authentication,” *BVAu 2010 - Bildverarbeitung in der Automation*, 1.

Ragavi, E. (2020) “Banknote Authentication Analysis Using Python K-Means Clustering,” *International Journal of Innovative Science and Research Technology* , 5(10).

Saluja, R. (2018) *Bank note authentication UCI Data*, *Kaggle*. Available at: https://www.kaggle.com/datasets/ritesaluja/bank-note-authentication-uci-data (Accessed: March 27, 2023).

Shahani, S., Jagiasi, A. and R., P. (2018) “Analysis of banknote authentication system using Machine Learning Techniques,” *International Journal of Computer Applications*, 179(20), pp. 22–26. Available at: https://doi.org/10.5120/ijca2018916343.

Sklearn (2023) *Sklearn.ensemble.gradientboostingregressor*, *scikit*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html (Accessed: March 27, 2023).

Waskom, M. (2021) “Seaborn: Statistical data visualization,” *Journal of Open Source Software*, 6(60), p. 3021. Available at: https://doi.org/10.21105/joss.03021.