# <https://shingi2020.github.io/tm_portfolio/portfolio.html>

### Unit 2 – Exploratory Data Analysis

Before any machine learning exercise can be undertaken on a dataset, there is need to understand what the data is and what it consists of. It’s important to note that as a data scientist, analysis doesn’t start with the dataset but with the problem statement. Meaning that one has to know what it is they want to answer and choose the right datasets to help them answer these questions.

The exploratory data analysis (EDA) process makes it easier for the data scientists to test their hypothesis, check assumptions, patterns and spot anomalies in the chosen data (IBM,2020). Through EDA, we can determine whether the statistical techniques being considered are appropriate, and if there are relationships between the variables.

During the group assignment, on Airbnb New York City listings, we were tasked with coming up with a business question and answering it using the listing data. By using EDA, we were able to formulate questions but in addition refine our questions if the exploration deemed our questions to redundant or gave inconclusive results or were outright not the right question to ask.

For the Airbnb EDA process, multivariate nongraphical and multivariate graphical were used to analyse the data. Giving a summary of the relationships between variables, and summary statistics. A similar exercise was performed on Wine Quality dataset presented in the e-portfolio. The task was to determine what characteristics are required for wine to be defined as being good quality. A correlation matrix was used to graphically show the strength of relationship between the variables.

In Unit 5, clustering was introduced which is an exploratory data analysis tool used to group objects in a manner that makes them similar to each other within the group than to other groups. This was applied to the Airbnb listing dataset. After having decided on the top neighbourhoods with the most listings and the highest priced listings. The goal was then to cluster the listings by highest priced and use other variables to create distinct groups which do not overlap with other listings in the other clusters. Lastly, we aimed to visualize these clusters geographically and to draw conclusions on whether the location or other listing characteristics played a factor in the listing price. The goal was to cluster highly priced listings in Brooklyn, Manhattan and Queens with similar characteristics. Using the elbow method, we were able to determine the number of clusters that needed to be created. From the analysis it was determined that the type of house had a bearing on whether clients were likely to spend more money.

### Unit 3-4

One of the core work areas of a data scientist is to be able to use machine learning to predict outcomes. We once again use the problem statement as our basis of undertaking a task of predicting outcomes. One such problem statement could be at what age clients are likely to purchase insurance. Logistic regression would be an ideal predictive modelling tool to use in this case. I would imagine a scenario where agents at a call center are given quotas to cold call potential clients. It would be of no use to call clients that fall out of the “likely to buy” age range. By being hyper focused, call center agents are likely to have more sales conversions, as opposed to aimlessly calling everyone on their customer database. The exercise was to create a predictive model that calculated whether a client would purchase insurance or not. From the results, clients below 30 years of age were less likely to want to purchase insurance.

### Unit 7-11

For these units, I found it easier to understand the logical workings by studying real world problems that use either Artificial Neural Networks (ANN) or Convolutional Neural Networks. An example of this would be customer churn prediction using ANN. For this example, the problem statement was how many customers were likely to switch to another mobile service provider by year end. Customer churn has direct impacts on revenue and profits, and the more customers leave the less revenue is brought into the business. Using ANN not only predicts the number of customers likely to leave but the type of customers and the reasons why they left.

The neural network consists of three layers, input, hidden and output layer. The dataset was split into 3 parts, training, validation and testing sets. The model was trained, and the performance evaluated for accuracy.

The end goal was to have a user input the necessary parameters and the model to predict whether a customer would remain a subscriber to the mobile service or not. Such insights are very useful for businesses and can be used as a basis to either change service offerings or create targeted sales and marketing strategies.

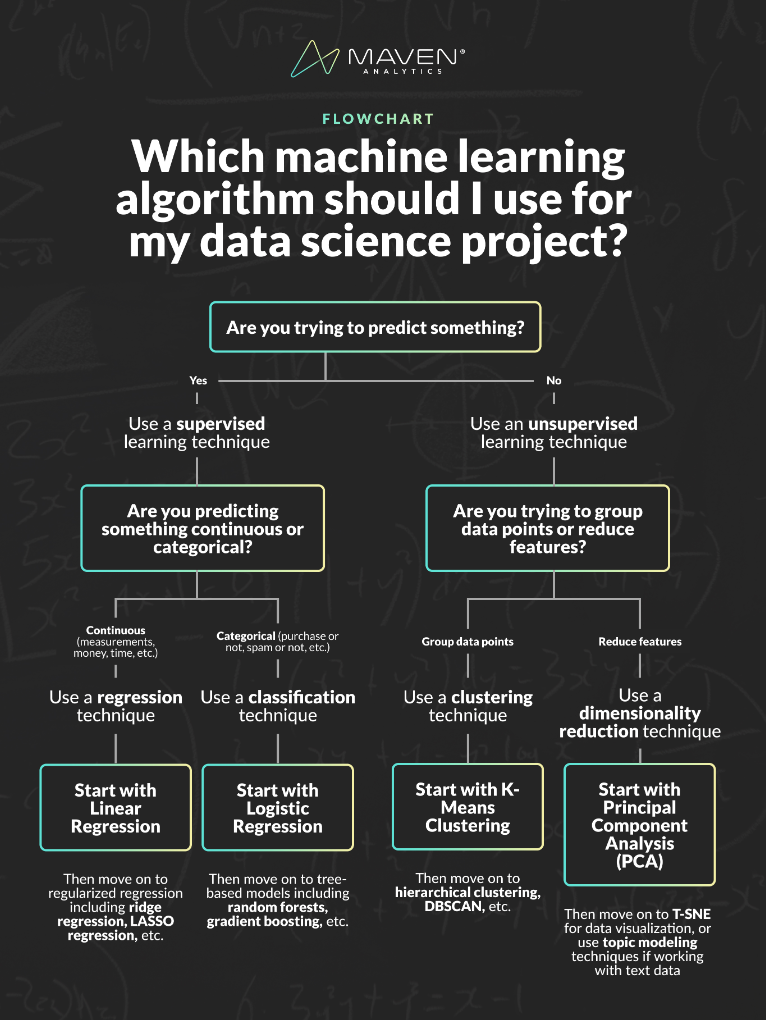
## Reflective Piece

This was my most anticipated module, due in part to its difficulty and complexity, which made me excited and scared at the same time. However, in the last 12 weeks, I have found It to be an enjoyable learning journey, with lessons and learnings that I have put into practice in my current work as a data analyst.

This module gave me a better understanding of how machine learning and data can be used in almost all everyday scenarios. From analysing Airbnb listings, to customer churn to building neural networks that can predict the correct image. Machine learning has become the standard in solving real-world problems, especially with the volume of data that is being generated daily. However, there are some legal, social and ethical issues that machine learning professionals must be cognizant of that have dire consequences for populations if not addressed. According to the Centre for Data Ethics and Innovation (Centre for Data Ethics, 2020), algorithm biases have amplified historic biases or created new forms of biases or unfairness, and there is a need to adopt more rigorous and proactive approaches to mitigating biases.

### Key Takeaways

To best explain how I understood the learning from this module, I will use this workflow developed by Maven analytics (Maven Analytics, 2023).



This flowchart summarizes the module on machine learning very concisely as it directs the user to a particular machine learning function to help answer the question posed. For example, as a data scientist I could be asked to predict whether customers will buy a product or not. Using this diagram and the units covered in the module, I would choose to use supervised learning techniques, and since the data is classified i.e. customer will stay or customer will leave, meaning this is either a “1” or “0” answer “Yes” or “No”, I would then select classification techniques such as logistic regression covered in unit 3.

### Group Assignment

The group assignment was a highlight of the course. Working alongside other students helped me gain a better understanding of the coursework. Furthermore, we encouraged each other as a group to explain our findings and code to each other during the group assignment. This in turn was a good way for me to test my knowledge and fundamental understanding of the theory behind the code I was writing.

The assignment also helped us understand and frame the business case of Airbnb and to think about the data to an end to grow the business and improve on sales and listings. One of the contributions I made was mapping the listings. The business analytics question which was posed was “What attracts customers in New York city to book premium listings”. Is it location? Amenities? Proximity to New York landmarks and shopping districts? Our approach to this question was to narrow down the most popular neighbourhoods by listing and then to cluster the premium bookings by price within each neighbourhood. As an added benefit, I suggested adding a map that shows these premium listings. The thinking behind this was that if it is visualised, we are able to see whether our hypothesis holds true. Are these listings in exclusive areas, or close to premium shopping areas or located in areas known to have apartments with spectacular views. The map view is available here: (<https://github.com/shingi2020/machine_learning/tree/main/Unit6_K_means_clustering>).

### e-Portfolio

For my e-portfolio, I expanded on my already existing portfolio that I have been working on. I created a branch in GitHub that covered the machine learning module exercises. On the website page of my e-portfolio I put together a summary of methodology, charts and reflections on each exercise. As part of the learning process, I undertook some exercises from Codebasics, Youtube page which was referenced in the Essex Lecturecasts.

During Unit 1 and 2, I contributed to one Wiki listing on Python Libraries for Machine Learning (https://www.my-course.co.uk/mod/glossary/view.php?id=985115). The article lists the 20 libraries that one should know if they are doing data science and machine learning. All the libraries used in this module were listed and a few others that weren’t touched but are worth knowing.

### Experiences

During the course of this module, I realised that all my past modules were leading up to this module. Each module from The Data professional, numerical analysis, deciphering big data, and data visualisation were all foundational to this module on machine learning. This made the module a lot more enjoyable as I could piece together each module and were it fits into the bigger picture. Further to this I found myself applying what I had learnt to my everyday work and trying to see how I could solve the current problems using machine learning and approaching data analysis problems in a systematic way.

Using what I have learnt, I want to use machine learning and data analysis in my line of work which is the humanitarian sector and predict the areas where climate change disasters are likely to occur and how many people will be in need of humanitarian aid. Using historical data on previous floods, and droughts that have impacted particular geographical locations, latest disaggregated census data on the current population figures, data collected by non-governmental agencies on vulnerabilities and aid given to the areas, I think I could be able to predict were aid efforts need to be focused on. I do believe that this would be something that would be impactful in my line of work and would remove the guess work from coming up with estimation figures of person in need.

### Action Plan

Having realised that each module in my PG Diploma in Data Science and Machine Learning is intricately connected. I realised that there were some gaps in my understanding of some critical theoretical elements.

That being said, I have tasked myself with reviewing my previous modules and getting additional foundational knowledge which is crucial to mastering machine learning.

The e-portfolio has become a show piece to show potential recruiters what I am capable of and that I am actively working towards mastering my data science skillsets. I will push myself to at least working on monthly data science challenges and uploading to the portfolio.

Finally, I acknowledge that I should have put more dedicated time aside to completing some of the e-portfolio activities. The field of data science is for someone who is into lifelong learning, and this is a commitment I have to adhere to in order to advance my career.

### References

1. IBM(2020). *What is exploratory data analysis(EDA)?.* Available at: <https://www.ibm.com/topics/exploratory-data-analysis> (Accessed: 19 July 2024).
2. Centre for Data Ethics (2020) *Review into bias of algorithmic decision-making*. Available at: <https://assets.publishing.service.gov.uk/media/60142096d3bf7f70ba377b20/Review_into_bias_in_algorithmic_decision-making.pdf> (Accessed: 20 July 2024).
3. Mavern Analytics, (2023). *Which Machine Learning Algorithm Should You Use for Your Data Science Project?* Available at: <https://www.youtube.com/watch?v=KcClHR3b1JA> (Accessed: 20 July 2024).