Assignment 4 Write-Up

1. **Case Study:**
   1. I decided to read the case study first without looking into the code first. This attempt was just to understand the flow and concepts used in the case.
   2. Once I was clear with the case and the technicality of it, I chose now to read the case study again with the code implementation of the techniques.
   3. Step 3 was to try to mimic the code on Jupyter notebook instance.
   4. I faced quite a few library errors, I chose to look for answers in the documentation of the library at first and then look for solutions online.
   5. Some of the libraries used in the code were no more in existence. I looked for libraries with methods that do exactly the same job.
   6. I made table of content and made sure to leave comments for easy readability.
2. **Machine Learning Notebook:**
   1. Here since the data wasn’t already given in the instructions, I went online looking for dataset that would best suit the purpose of the notebook. I generally use datasets provided by UCI in their Machine Learning Repository as they have good documentation and quality data.
   2. Once I decided I wanted to work on UCI’s car model specification data. I imported the data to my notebook using their Url. I avoid importing data after downloading locally to make it easier for submission and grading.
   3. After getting the data, I did a quick data quality check. Data had missing values, missing values above a certain percentage led to dropping 1 column, for other values in rows Imputation methods were used to replace them with a median value.
   4. The final outcome of the data cleaning process was a better quality data. Since I was planning to run classification models on the data and data had quite a lot of string values, my next step was to using few encoding methods to convert these columns to suitable integer or categorical values.
   5. Once I was through with the pre processing. I had to make target classes from price column which was continuous in nature. I decided to break it into two classes “Affordable” and “Expensive” taking mean price as a cutting line between them both.
   6. Further, I split the data into 30/70 portion and used various classification methods on the data.
   7. I used evaluation scores to look for the best performing model of them all. I made sure to take Affordable class as a positive outcome.
   8. Later, I tried cross validation methods along with regression methods to boost the performance of the model.
   9. Finally, I made sure to build a pipeline that does above mentioned Machine learning steps automatically.
   10. Yet again making a Table of content for easy moving though the notebook was made and extra comments were added for better understanding.