Saivee Phatak, 210252
OPSM322

Machine Learning I – Final Project

For the final project, I have divided my objective into three codes:

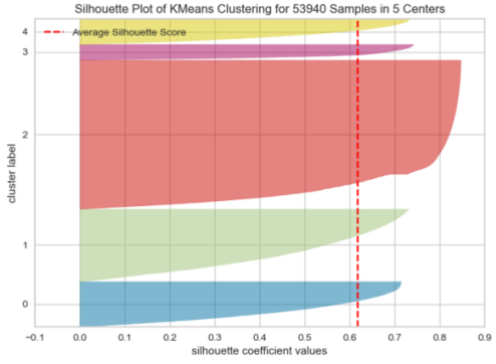
Code 1: This code deals with unsupervised learning, and includes the code for K-means Clustering, Hierarchical Clustering and Spectral Clustering.

Code 2: This code deals with supervised learning- classification, and includes the code for LDA, QDA, Mixed NB, DT and Random Forest.

Code 3: This code deals with supervised learning- regression and includes the code for Linear regression, Ridge, and LASSO +elastic net regularisation, KNN regressor and Random Forest Regressor.

Code 1- Unsupervised learning on Diamond Dataset

- The objective for unsupervised learning for the diamond dataset is to cluster the diamonds based on various factors, including the likes of clarity, depth, height, width, colour, and other factors.
- For clustering of diamonds, I did K-means Clustering, Hierarchical Clustering and Spectral Clustering.
- I created dummy variables out of clarity and colour, which is required for all the clustering models. I also normalised the dataset for a better result, since it did not follow normal distribution.
- I ran the Elbow plot and the KElbowVisualiser, and I got the inflection point as 5. Hence, I decided to run K-means with 5 clusters.
- The normalisation reduced the WCSS from 57125590025.26 to 49569.33, showing that normalisation was helpful.
- For K-means Clustering, I got a silhouette score of 0.3180, and the following clusters:



- On profiling the clusters (profiling is in the code as comments), the clarity of the diamond seems to be the deciding factor when it comes to clustering.
- Due to computational limitations, I could not run hierarchical and spectral clustering on the entire data. Hence, I decided to run it on only 10% of the data in order to at least get an idea of what the models might look like.
- For hierarchical clustering, I got a silhouette score of 0.2556 for 6 clusters using ward linkage.
- For spectral clustering I got a silhouette score of 0.2681 for 6 clusters.
- Ultimately however, **K-means clustering** provided me with the best silhouette score, and this is the model I would use on this dataset.

Code 2- Supervised learning on Diamond Dataset- Classification

- For my Code 2, I decided to use the same diamond dataset. I added the dependent variable, which was the 'cut' of the diamond which was categories like premium, fair, average etc, and all other variables including clarity, depth, height, width, colour, and other factors as my independent variables.
- I ran LDA, QDA, Mixed Naïve Bayes, Decision Tree, and Random Forest on this dataset. I decided to cut out SVM due to computational limitations.
- Following were my accuracies for all the models that I tried:

	Original Model	Grid CV search
LDA	60.7712	61.7445
QDA	57.8605	64.7107
Mixed NB	57.3785	57.4898
Decision Tree	70.9306	76.6036
Random Forest	76.6499	76.585

- As seen above, Grid Search for the best parameters improved almost each and every one of the models, and it remained the same in the case of random forest.
- In this, through the correlation matrix and F1 score (given in the code), ultimately decision tree comes out as the best model, edging out random forest.
- LDA and QDA on the other hand, performed just average, which could be attributed to the fact that the data is not normally distributed.

Code 3- Supervised learning on Insurance Dataset- Regression

- For my Code 3, I decided to use an insurance claim dataset, which I primarily dealt with using regression. The insurance claim dataset's objective was to predict how

much money an individual would get from an insurance company based on various factors such as age, blood pressure, BMI, no. of children etc.

- For this dataset, I ran linear regression, Ridge and Lasso regression, Elastic Net, KNN Regressor and Random Forest Regressor.
- Following is the RMSE and R squared results for all the models that I ran:

	RMSE	R^2
Linear	6522.43	0.703786
Ridge	6526.21	0.703443
Lasso	6539.84	0.7022
Elastic Net	6526.74	0.70339
KNN	8927.46	0.44506
Random Forest	4957.27	0.82889

- As seen according to the results, Random Forest is the best regression model to run on this dataset. It provides the lowest RMSE, and highest R square out of all the models.
- Alongside RF, Elastic Net is the next best model for the dataset.
- The Lasso model does not work particularly well, since when it tries to do feature selection based on multicollinearity, it ends up removing some variables that even though are highly multicollinear, are quite important to the model.
- The KNN regressor performed below average, despite trying to find the best parameters for the same.

There were several challenges I faced during the project, including computational limitation. The datasets were not ideal, and hence interpreting the clusters for the same required a bit of research. I had to make several tweaks to the code in order to allow it to fully adapt to the dataset, which also took a bit of external research.
Overall, I feel like I gained a clear insight into the different types of models, their functionalities, the strengths, and weakness of each model, and how to best optimise them. My datasets provided some promising results that I will continue to work on even beyond this course.

My final video presentation is submitted in the file itself. However, if it does not open, here is a link:
https://drive.google.com/drive/folders/1yj9op2zMLi6y-m3UXcrSUulDKOT6uWOY?usp=sharing