CSE 578: DATA VISUALIZATION

Individual Contribution Report

Aditya Chayapathy (1213050538)

Reflection:

Role:
Throughout the course of the project, I was assigned the role of a data scientist.

Responsibilities:

1. Understanding the dataset.
2. Univariate and multivariate analysis of the following set of features with help of data exploration techniques such as statistical analysis and data visualization:
    a. Capital-gain
    b. Capital-loss
    c. Sex
    d. Hours-per-week
    e. Native-country
3. Machine learning analysis
4. Documentation

Lessons Learnt:

Some of the important lessons learnt during the course of the project are as follows:

1. It is of utmost importance to understand the dataset thoroughly before performing any analysis on it. This involves understanding the meaning of each of the features and their correlation towards the class labels. Some of these relationships are intuitive while others require deeper digging.
2. For univariate feature analysis, it is beneficial to automate the process of performing statistical analysis and plotting the corresponding visualizations. With that, one can invest more time in analyzing the results.
3. Not all statistical measures or visualizations reveal patterns and hence one should patiently analyze all the plots.
4. For continuous data, it would be beneficial to understanding the underlying distribution to make better judgements.
5. For categorical data, analyzing the class label distribution for each unique value can help in judging how important that category would be in the classification problem.

6. For multivariate analysis, it is important to figure out which combination of the features reveal patterns. This requires univariate analysis of each of the features and understanding the effect of each of the features towards distinguishing the class labels.
7. Machine learning analysis should be done only when one has a good domain knowledge of the underlying dataset.
8. Machine learning analysis should be done in the following order of steps:
    a. Understanding the problem statement
    b. Data cleaning: The data set may contain incomplete or missing values. It is important to clean the dataset by either removing or replacing missing values to ensure consistency throughout.
    c. Feature engineering: It is important to identify the set of features that influence the decision of class prediction. Not all features may contribute towards classification. Also, it is important to convert categorical data into numerical data as most of the libraries work with only numerical values. While converting, it is important to ensure the meaning and contribution of the categorical values are preserved.
    d. Data sampling: It is important to ensure that the training dataset has data with respect to all classes in an approximate 1:1 ratio. Otherwise, the ML algorithms tend to be biased towards one or more classes.
    e. Data transformation: It is important to normalize the data along each of the features so that equal importance is given to each of the features by the ML algorithms.
    f. Identifying ML algorithms to be used
    g. Training ML algorithms on the training dataset
    h. Hyperparameter tuning of the algorithms
    i. Evaluating the performance of the ML models on the test dataset
    j. Operationalizing the ML models

Assessment/Grading:

Personally, this project has been a great learning experience for me. Having had minimal prior experience in the field of data science, through this project I have had the opportunity to expand my skill set and improve my data exploration capabilities. Our team followed an agile methodology with which the roles and responsibilities that were assigned to individuals were made very clear from the very beginning. Each of us contributed equally towards achieving the common goal and were able to complete the assigned tasks well within the stipulated timeline. We worked together as a team and were ready to lend a helping hand when needed from other team members. Having been responsible for the machine learning analysis, I have made a huge leap towards improving my skills in this field. Data visualization is a great tool and with this project, I've had the opportunity to use it to its complete potential. To conclude, my individual assessment of the project would be 10/10. I'm very satisfied with my contribution and the effort of the rest of the team.

Future Application:

The following are some of the major skills that were acquired during the course of the completion of the project:

1. Data visualization using python
2. Univariate feature analysis
3. Multivariate feature analysis
4. Statistical feature analysis
5. Feature engineering
6. Machine learning analysis using python

With these skills, I now have a strong foundation in the space of data analysis and machine learning that can be applied to all my future projects.