Overall Presentaion and Reflective Report

Process of solving problems and learning to use notebooks

Being from the coding background, I was familiar with the Jupyter notebook and how to use it. Using the machine learning models and wrapping my head around new concepts seemed pretty intimidating at first. As the weeks went by, I became more confident with the practice from the workshops and lectures.

Stating the major learning which were the place of exploratory data analysis (EDA) and data preprocessing. As such, I mastered the skill of working with missing values, converting categorical variables to suitable formats, and dividing the data into training and testing sets, which are the type of pre-processing activities that have to be done before an actual machine learning model building activity. Everywhere in the portfolio, I utilized different algorithms and techniques, making use of linear regression as a basis and moving to slightly more elaborate forms like polynomial regression, random forest, and the gradient boosting. This procedure was beneficial in the sense that I was in a position to discover the strong and weak points of each of the approaches and their dependence on problem types. Speaking of evaluation metrics I learned various of them i. e mean squared error (MSE), mean absolute error (MAE), R-squared, accuracy, F1-score, and area under the ROC curve (AUC). Employing these metrics allowed me analyze my model's performance as well to pick the techniques that worked on the dataset.

All in all, the learning process of solving problems with Jupyter Notebooks has been a great learning experience. I have acquired hands-on skills in data-analysis, machine-learning and coding that will be useful in future undertakings and different research endeavors.

Future Interests

Moving forward, I am overwhelmed with the anticipation and eagerness to witness the far-reaching possibilities of data science. The industry is in a state of transformation which brings forth new challenges and opens up a new pathway. Another aspect that excites me is the crossroad of data science and machine learning in which I can apply advanced forms of algorithms and models to discover hidden patterns and predictions.

Moreover, I also have a keen interest in the application of data science methods to real-world issues, including business, healthcare, or environmental sustainability areas. When data is utilized to its full extent, I think we can facilitate innovation, process improvement, and better decision making which in turn have a positive impact on our society.

To sum up, this unit in data science was undoubtedly impactful, as I acquired an important skills, developed a problem-solving attitude and became passionate about data-based explorations. While I acknowledge that the road ahead will not be easy, I am confident that with the knowledge and expertise I have gained so far, I will not only be able to tackle the challenges ahead but also join the field of data science in its continued evolution.

Choice of Dataset in the Portfolio

"StudentsPerformance.csv" was used as part of the portfolio. Dataset is about the Students Performance in Exams. It contains the marks secured by the students in various subjects. "StudentsPerformance.csv" dataset was selected because it shows the interesting problem of predicting the student's performance by using different features that can be used for the education research or education policy-making.

Problem Identification

The first problem dealt with in this portfolio was to develop predictive models that could determine a student test score based on the presented features: 'gender', 'race/ethnicity', 'parental level of education', 'lunch','test preparation course', 'math score', 'reading score','writing score'. Through this understanding, educators and policy makers will be able to make decisions that are well informed, and these, in turn are intended to improve educational policy making.

Choice of Machine Learning Models

A selection of machine learning models in my portfolio were random forest regression, gradient boosting regression, polynomial regression, and linear regression. Random forest regression and gradient boosting regression are the ensembles that aggregate multiple decision trees and usually show better predictive precision and robustness as compared to individual models. Linear regression gave a simple reference model in order to see the linear relations between the features and the objective variable. However, polynomial regression was applied such that the performance of the model could be improved because of non-linear relationships.

All these models were applicable to the problem variations where the target variable was continuous and the features included both numerical and categorical types.

Insights and Conclusion drawn from the study

As a result of the data analysis and planning, series of insights and conclusions are obtained. Polynomial regression model has shown better results over that of a linear regression model, thereby indicating the existence of non-linear relations between features and the target variable. I used exploratory data analysis, data preprocessing, and applied various machine learning models like linear regression, polynomial regression, random forest, and gradient boosting. I gained expertise in handling missing data, encoding categorical variables, splitting data, and evaluating models with metrics like MSE, MAE, R-squared, accuracy, F1-score, and AUC. The process allowed me to understand the strengths, weaknesses, and applicability of different techniques to different problem types. Overall, this learning experience equipped me with valuable data analysis, machine learning, and coding skills applicable to future research endeavors. Moving forward, I am eager to explore advanced algorithms, real-world applications across industries, and the transformative potential of data science and machine learning.