APEX FIND - DATA MINING ASSIGNMENT WRITE UP

PERFORMED PROCEDURES

1. Generated statistical information about the different subject scores to get a general idea regarding student performances
2. Calculated the average score of each student
3. Generated a heatmap for each subject score and average score to check for any correlations between the subject scores and average score
4. Generated a box plot for the following
    - Lunch type vs student scores
    - Parental level of education vs student scores
    - Test preparation course vs student scores
    - Race/ethnicity vs student scores
    - Sex vs student scores

5. Changed all categorical variables into dummies during the data processing steps in order to run the Random forest regressor model.
6. Created a train and test set and fitted the sets with a ridge regressor model.
7. Calculated the mean squared error score after fitting the ridge model.
8. Used the random forest model classifier to identify feature importances in order to find out which categorical variable has the greatest role/factor in student scores.

FINDINGS

1. The students scored better grades in writing and reading compared to maths on an average. Since the standard deviation and interquartile ranges of grades in all three subjects were similar it shows that the grades are rather evenly distributed within a range.

2. The heatmap points out that there is a higher correlation between the reading and writing scores of a student compared to that between maths and reading/writing. This indicates that students who score well in reading tend to score well in writing and vice versa but the statement does not hold true for when maths is concerned. Students who score well in reading tend to score better in maths as maths and reading have a higher correlation than what maths and writing have.

3. There is also less of a correlation between the average score and the maths scores compared to the average scores and the reading/writing scores. This seems logically correct given point (2).

4. From the boxplots it is seen that students who had the standard lunch scored better in all three subjects and hence also had a better average compared to the students who had the free/reduced lunch. This makes more sense as science has always backed the correlation that better and nutritious food has with that of brain function, concentration and mood. Nutritious foods provide the body and mind with the energy needed to grow, feel well, be active, stay healthy and learn.

5. Unsurprisingly, childrens with parents that have master's degrees score higher on average in all subjects when compared to other levels of education. Additionally, the minimum score for that group is higher in all subjects when compared to all other groups. Higher levels of education may mean a higher socioeconomic status, as well as a greater emphasis placed on the importance of education, leading to higher test scores.

6. An unexpected find is that parents that fall into the "some high school" had children that scored higher on average than the "high school" group. This seems to fly in the face of what this data has indicated, which is that children with parents that have more resources and place a greater emphasis on education tend to have higher test scores. It is difficult to know what this exactly means, this variance could just be a result of a small sample size, as there were only 537 subjects in this group. One possible explanation could be that parents at the lowest level of education may want better for their own children, which causes them to imprint a heavier emphasis on education.

7. The boxplots indicate that the students who completed or took the test preparation course scored better grades in all three subjects and hence also had a better average than the students who did not take the test preparation course. The stripplot indicates that there are a few students who did not take the test preparation course and still ended up scoring very well. The strippoyt also indicates that the lowest grades were scored by students who didn't take the test preparation course. Overall the test preparation course is helpful for students but the course may be subjective for certain students.

8. The score of students depends highly on which race/ethnicity and family background they come from. We see from the boxplot that the students from group E score the highest average grades and students from group A score the lowest average grades out of all groups. Students from group B, D and C score very similar grades in all three subjects but students from group D have a slightly higher average compared to the students from group B and C.

9. It is seen that boys on average score higher than girls in maths but girls score better in reading and writing. Girls also have a better overall average compared to boys.

10. From the feature importance plot we can see that the test preparation course and the type of lunch that a student has are the biggest factors in determining what score he/she gets. Parental level of education and student race/ethnicity are factors that have lesser importance on what scores a student gets. Parents who attained a bachelor's degree and students who come from group B are the two most influential out of those categories towards what grade a student gets. Gender of a student does not have a massive importance in determining test scores.

CONCLUSION

The school should take the following steps in order to increase student scores -

- Ensure all students eat the standard lunch. The school should subsidise the standard lunch for the students who can't afford it.
- Promote all students to complete the test preparation course.
- Encourage students to focus on highly correlated subjects such as reading and writing. Preferably reading as it also has a better correlation with maths compared to writing.
- Encourage/help boys to focus more on reading and writing and do the same with girls for maths.
- Provide extra resources to students who’s parents have a level of education of “high school” and “some high school” and to students who come from group A. These students are prone to not have the best resources in order to score better grades.

The steps above are listed in order of decreasing importance/urgency in order to improve student scores.