Near-Earth objects Report (Team 5)

This report presents an analysis of near-Earth objects (NEOs) to predict whether or not they are dangerous and which variables are significant indicators of hazardous objects. The study used both linear regression and classification models to achieve the objective. The dataset used is a CSV file named "neo.csv" and contains data on NEOs, including their orbital and physical characteristics. The data has 10 variables, including id, name, est_diameter_min, est_diameter_max, relative_velocity, miss_distance, orbiting_body, sentry_object, absolute_magnitude, and hazardous.

The study used multiple linear regression to identify the significant predictors of hazardous NEOs. The initial model included all the predictors, and the model with the highest R-squared value was chosen as the best model. The final model included miss_distance as a response variable and only two predictors: relative_velocity and absolute_magnitude. These predictors were significant, and the model's R-squared value was 0.13, which suggests that 13% of the variance in the response variable can be explained by these predictors. As we developed and trained the linear regression model, we removed the predictors est_diamter_min and est_diameter_max. These predictors did not influence the model in a significant way, and they both gave the model a high multi-collinearity.

The study also used diagnostic plots to check the assumptions of the linear regression models, including Residuals vs. Fitted plot, Normal Q-Q plot, and Scale-Location plot. The Residuals vs. Fitted plot showed somewhat equally spread points in a downward linear pattern, the Normal Q-Q plot showed a pattern that slopes upward towards a line, follows a linear path, then deviates horizontally, and the Scale-Location plot showed a high concentration of residuals in the middle of the line and gradually thinner spread towards the beginning and end. The study also tested for multicollinearity using the Variance Inflation Factor (VIF) and found that there was very low multicollinearity in the final model.

In the the logistic regression model in the study is the hazardous01 variable, a binary variable indicating whether or not an NEO is hazardous (1 for hazardous, 0 for non-hazardous). The predictors in the study include 5 other variables such as est_diameter_max, est_diameter_min, relative_velocity, and absolute_magnitude, and miss_distance. My selection is to randomly split the available data into a training set and a test set. The ratio of the split is to use a 80/20.

The logistic regression model was used to classify the NEOs as hazardous or non-hazardous. The study created a binary variable named hazardous01 to use as the response variable in the model. The logistic regression model achieved an accuracy of 98.75% on the test dataset, indicating that the model can predict whether or not an NEO is hazardous with high accuracy.

In conclusion, this study has shown that it is possible to predict whether or not NEOs are hazardous using linear regression and classification models. The significant predictors of hazardous NEOs are relative_velocity and absolute_magnitude. The logistic regression model achieved a high accuracy of 98.75% on the test dataset, indicating that the model is suitable for predicting hazardous NEOs.