Milestone 2 Report

Group 20- Zaid Hasan, Aditya Jain, Swate Patel, Adit Nuwal, Roman Marach.

Report on Multiple Regression Model for Predicting Insurance Charges

A multiple regression model is developed and reported in this report in order to account for and predict the insurance charges using a few selected features from the dataset. This model employed the features such as:

• Age: The age of the individual that might be related to the expected cost of healthcare services as elderly individuals are likely to incur greater health expenses.

• Sex: A variable that captures information about an individual's gender.

• BMI: Body Mass Index (BMI) is a measure of body fat for adults that is calculated based on height and weight. BMI is commonly associated with specific health effects.

• Children: The number of dependents in the form of insured children under the policy which is likely to contribute towards total charges.

• Smoker: A variable showing the smoking status of the individual which is usually associated with increased health risks.

• Region: The geographical factor, implemented by using dummy variables for regions, which are the areas for which there could be a difference in costs.

These features were considered for the prediction of the target variable, Insurance Charges as these were found to be closely related to the particular health care and lifestyle characteristics which usually determines the health expenditures of the individual.

Model Implementation and Training

The collected dataset is divided into subsets, with 80% allocated for training and 20% for testing purposes. With the presence of categorical variables like "Sex," "Smoker," and "Region," one-hot encoding was applied for preparing the data for the regression model. The regression model was developed using ordinary least squares (OLS) estimation, in which the np.linalg.inv function helped to derive the regression coefficients.

Model Evaluation

Two important measures of the model were utilized to determine its performance as highlighted below:

• Mean Squared Error (MSE): The data collected from the model's test output gave a MSE of close to 37,175,951.41, which provides an average of squared difference between the forecaster's estimates and the actual insurance charges and so its values are expected ranges of the actual number.

• R-squared (R^2): From the model, it was observed that the R^2 from the model was 0.758 meaning that 75.8% of the variance that is presented in the insurance charges is described/ predicted by the variables of the model.

Interpretation of Results:

The R^2 score of 0.758 surpasses the prescribed acceptable limit of 0.7 and further suggests that the model can explain a great portion of the spread in the insurance charges. The MSE in this case being considerably high, is understood to represent a model's accuracy in comparison to other models. Exploring the ranges for MSE that would be suitable to the data set would deepen the understanding of the level of accuracy of the model.

Feature Impact Analysis

A review of the regression coefficients led to the identification of the factors that were considered to contribute the most influence on insurance charges:

• Smoker: The fact that one was a smoker increased the charges for insurance by a great margin most probably due to the health threats posed by smoking.

• BMI: It has been determined that the insurance charges tend to be higher for clients with a high BMI as high BMI can lead to a number of health factors.

• Age: insurance charges are strongly related to age. This can be explained by the fact that older people tend to spend a lot more on medical expenses.

Conclusion

Insurance charges can be predicted by use of the multiple regression model where R2 score achieved was 0.758 which can be said to be better than the average expectations. This means that the model explains high percentage of the variation in the charges although there was still some room for more changes. Subsequent changes may include testing of models that are relatively more sophisticated or further feature engineering so that the forecasting accuracy is higher especially when lowering MSE is being the goal.