2022

Logistic Regression – Graded Assignment

LEAD SCORING – LOGISTIC REGRESSION - ASSIGNMENT

BHAVESHKUMAR THAKER

UPGRAD – CHANDIGARH UNIVERSITY – M.SC. IN DATA SCIENCE

Summary Report

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, their profile, the time they spend on site, how they reached the site and the conversion rate.

Following steps were followed to gather the learning from the assignment:

- **Load and Clean the Data**

  Many features in the data, had null values or option "select". Features having more than 40% missing data were dropped. To not to loose data, many null values were replaced by mode or median of the feature.

- **Exploratory Data Analysis (EDA)**

  Data analysis was done for target variable, categorical variables, numerical variables, and their observations were noted. Bivariate analyses were done to understand the relationship between features.

- **Data Preparation**

  Outlier analysis were done and removed the outlier values. Converted binary features into zeros and ones while for categorical features, dummy variables (one hot encoding) was performed.
  Dataset correlation was checked for highly correlated features and dropped features with very high correlation.
  Spliced the data into train and test sets and applied MixMaxScaler to normalize the data.

- **Model Building using Logistic Regression and Recursive Feature Elimination (RFE)**

  GridSearchCV and RFE was used to identify top relevant features. Then logistic regression was applied on the RFE selected features. Plotted ROC-Curve and precision-recall trade off to find the optimal cut-off point for predictions.

- **Model Building using Statsmodel API and Variance Inflation Factor (VIF)**

  VIF was used to drop high VIF features. Then statsmodel API for logistic regression was applied on the VIF selected features. Plotted ROC-Curve and precision-recall trade off to find the optimal cut-off point for predictions.

- **Model Building using RandomForest Classifier**

RandomForest Classifier was applied on all the features of the dataset and trained the model. Plotted ROC-Curve and precision-recall trade off to find the optimal cut-off point for predictions.

- **Model Evaluation**

  Applied the model(s) on test set and a confusion matrix and a classification report were created to understand the model's accuracy of prediction on unseen data.
  Optimal cut-off point using accuracy, sensitivity, specificity from the graph was 0.38 while optimal cut-off point using precision-recall from the graph was 0.445.