2. Data Preprocessing

2.1 Data Description

Our data set was collected and published by the Stack Overflow website and is also the full, cleaned results of the 2018 Stack Overflow Developer Survey. Survey included 129 questions, most of which are multiple choices Questions cover education, employment, coding experience & habits, demographic information, attitudes towards different topics and Stack Overflow product feedback and review. All the respondents were recruited primarily through channels owned by Stack Overflow. In the released version, there are 98,855 responses in total.

2.2 Data Cleaning

2.2.1 Delete irrelevant questions

Because our problem of interest is whether coders/developers with various backgrounds have different attitudes towards AI, there are many questions not related to the problem of interest in the original survey, such as attitudes towards advertisements on Stack Overflow, activity or engagement on Stack Overflow, product feedback, etc. Therefore, we screened all the questions in the survey and manually deleted those irrelevant questions in order to prevent interference with the later model. After deleting, there are still 58 questions remaining.

2.2.2 Convert categorical variables to binary dummy variables

For continuous variables in the data set, we can directly use them in modelling, however, for categorical variables, we are supposed to use reference cell coding method to convert them to binary dummy variables. For example, in terms of the variable Hobby, there are two different categories: Yes and No. Therefore, we choose the No category to be the reference group and convert two categories into dummy variables 1 and 0. In the data set, we also make some special transformations of the categorical variables. For instance, because the satisfaction variables assume that differences between adjacent groups are equal and there is an increasing trend under sequential categories, we convert job and career satisfaction into ordinal integer scores (1: extremely dissatisfied, 7: extremely satisfied). In addition, we decide to convert country to continent using country code because there are too many types of country variables.

2.2.3 Fill in missing values

After prescreening, there is only 3846 complete responses out of 98,855 responses. If we simply drop all the cases with NA, we will lose a substantial amount of information. The following steps are how we deal with NAs. Firstly, we drop responses with more than 50% unanswered questions. We regard the responses with more than half of the questions unanswered as invalid responses. Then, we use the KNN method to impute all the missing values. After imputation, we obtain a feature matrix of 69,083 samples and 357 independent variables.