EDA

Beibei Du, Wafiakmal Miftah, Suzanna Thompson, Alisa Tian

2022-11-04

IDS702 Final Project Plan Part 2

Overview

- The data used in this project comes from Kaggle with 607 observation with 11 variables. This dataset does not null or missing value. This is the link to original post: Data Science Job Salaries
- The Research Questions that we come up with are:

1) Which factors are associated with an increase in salary for data science jobs? (Continuous outcome)
2) How do company size, company location, employment type, employee residence and job title affected the remote work ratio of a data scientist? (Discrete outcome) For this question, there are three possible values for the remote work ratio, 0, 50, and 100; these signify an in-person job, a hybrid job, and a fully remote job, respectively.

Models

We will perform different models on the two research questions that we have separately:

1) For the first question, to examine the changes in Salary (USD), a continuous variable as our outcome/response variable, we are considering to use Linear Regression Models, more specifically, Multiple Linear Regression Models. We will fit and assess various models considering experience level, employment type, employee residence, remote ratio, job title, company location and company size as variables. To find the best model in predicting Salary (USD), we will do the selections method mentioned in variable selection.
2) For the second question, the response variable (remote ratio) is a categorical variable with 3 levels: 0, 50% and 100%, which represent an in-person job, a hybrid job, and a fully remote job, respectively. Thus the multiple linear regression model from first question is not applicable here. In this case, we can use multinomial logistic regression model. The log odds of remote ration will be calculated as a combination of all the predictor variables we are interested in (such as the company size, percentage of supporting remote job, employment type).

Variable selection

1) For the first research question, since we are taking multi-linear regression model as our base model. We will use Forward, backward,stepwise selections to pick the best features. Besides, taking AIC,BIC, Adjusted-R-Squared into consideration is highly valued as well. Since forwad and backward selections have their own disadvantages and stepwise is the main selection we are taking into consideration.
2) For the second research question, our response variable is a categorical variable with 3 levels. Thus to select features from multinomial logistic regression model, we will use chi-squared test and change in deviance to select the best model.

Challenges

A challenge we'll need to address before modeling is the amount of job titles that are present in our data. To solve this, we will collaspe certain job titles into relevant categories. In that, if the job title "NLP Engineer" only shows up once, we will combine this with a similar job title or exclude it on the whole. This decision will be informed by domain knowledge and field experts. We face a similar problem with employee_residence and company_location. Not only is more than half of our data from the United States, Great Britian, and Canada, but the two variables are also highly correlated. Because we're investigating the relationship between work place attributes and the remote-ratio, we will include both location-based variables. There are various common data issues that need to be addressed before modeling. We consider them below. Messy data is defined to be a dataset where values are unstandardized, unorganized, or bias. Largely, our data is clean. The only potential messiness in our data comes from the large number of job titles as mentioned above. Another common data issue is a lack of data. As our data has 607 entries, we do not face this problem. Finally, confounding variables can cause modeling issues as there may be variables that are related to our questions at hand that are not present in our data. Because our research questions are somewhat poignant in their phrasing, we will not encounter this as a data issue. One might suppose that a level of education may have an effect on a data scientist's salary as it tends to have an effect in other fields. However, because data science is a relatively new field, there are not much data on whether or not a higher-education degree has an effect on the ensuing salary. For this reason, we find no issue in assuming it has marginal effect.