EDA

Beibei Du, Wafiakmal Miftah, Suzanna Thompson, Alisa Tian

2022-11-01

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, our output/response variable is a continuous variable. Thus we first consider the Linear Regression Models. We assume the [`which variables???`] are the ones that might be potential ones that might increase the salary of data scientists. 2. For the second question, the response variable is a categorical variable with 3 levels. Thus the model choice from first question is not applicable here. We can use logistic regression in this case.

Variable selection

1) Forward, backward,stepwise selection; AIC,BIC, Adjusted-R-Squared
2) Chi-squared test and change in deviance

Challenges

TBH