

Final Project:
Data-related Job Salary Prediction

Background Introduction

In the past few years, more and more companies have needed data analysts, and many universities have also opened majors in business data analysis. Many tech companies have been incentivizing job seekers with staggering salaries to encourage more people to join the industry. But since the new coronavirus ravaged the world in 2020, the global economy has begun to fluctuate wildly, and the labor market has also received a considerable impact. To snatch high-end technical talents, many companies offer salaries beyond the salary range of fresh graduates to attract talents, which leads to very unhealthy and vicious competition. Therefore, our project hopes to analyze salary data and make reasonable predictions, to help major companies effectively control their labor costs and attract talents.

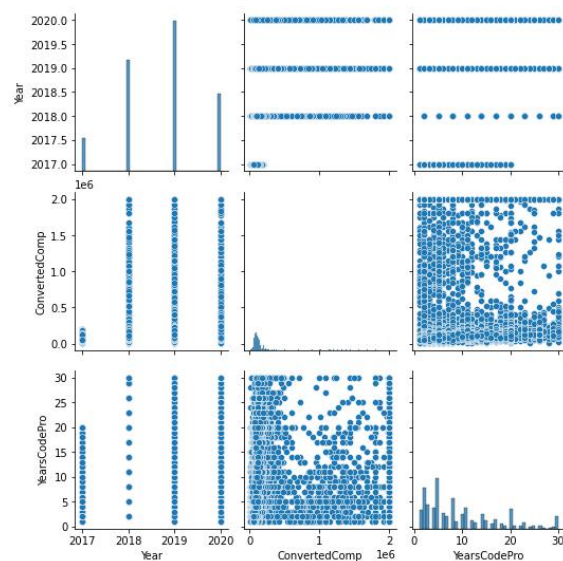
Data Description and Clearing

The raw data for this project comes from Stack Overflow's annual developer survey. Most respondents work in data-related jobs (data scientists, machine learning specialists, database administrators, business analysts, data engineers, etc.). The respondents' resources fill in the data. Judging from the final dataset compiled from all the questionnaires collected at the end, it has 15 variables and 33.6k results. A detailed description of this data lists above:

Variables	Data Type	Description
Year	Numeric	Date
Hobbyist	Categorical	Is programming your hobby?
ConvertedComp	Numeric	Convert salary in U.S. Dollars
Country	Categorical	Which country does data analyst come from?
DatabaseDesireNextYear	Categorical	Database environments interviewees have done extensive development work
DatabaseWorkedWith	Categorical	Database environments interviewees have done extensive development work
DevType	Categorical	Job title

Edlevel	Categorical	Which education level is a data analyst?
Employment	Categorical	Part-time or full-time?
JobSat	Numerical	job satisfaction from 0 to 10
LanguageDesireNextYear	Categorical	Programming language
LanguageWorkedWith	Categorical	Programming language
OrgSize	Numerical	How many employees does the company have?
UndergradMajor	Categorical	What is your major?
YearsCodePro	Numerical	How long do you start to code?

For exploratory analysis, my project uses the pair plot to find the correlation between variables. As we can see from the graph below, since most parts of our datasets are categorical data, it is difficult for following correlation analysis to figure out in-depth. But as we can see from the current graph, the most relevant variable for the variable we want to predict is the variable YearsCodePro. So, in the next step, we need to clean our entire data set to analyze the relationship between each variable and annual income deeply.



In terms of data processing, observations with many null values in the dataset are deleted in the first step. Specific breakdowns of some variables in the dataset (for example, we of undergraduate majors are unified, and the respondents' job satisfaction is unified based on the Likert Scale - 9 most satisfied, one most unsatisfied) are unified, which is the foundation of the following analysis. Then, since the object of our project is the analysis of wages in the United States. According to the research data released by the US federal government in 2020, the minimum wage in the United States 2020 is \$13,920, so it is necessary to delete the observations, including non-US results, that wages less than 13920. Also, converting some variables into more meaningful variables (Data scientist or machine learning specialist, Database administrator, Data or business analyst, Engineer, data) is crucial due to the ambiguity of several observations in the dataset (DevType, DatabaseDesireNextYear, etc.).

After all the variables are converted into the format we need, the next step is to convert the categorical data existing in the dataset into dummy variables, thanks to modeling requirements.

Modeling

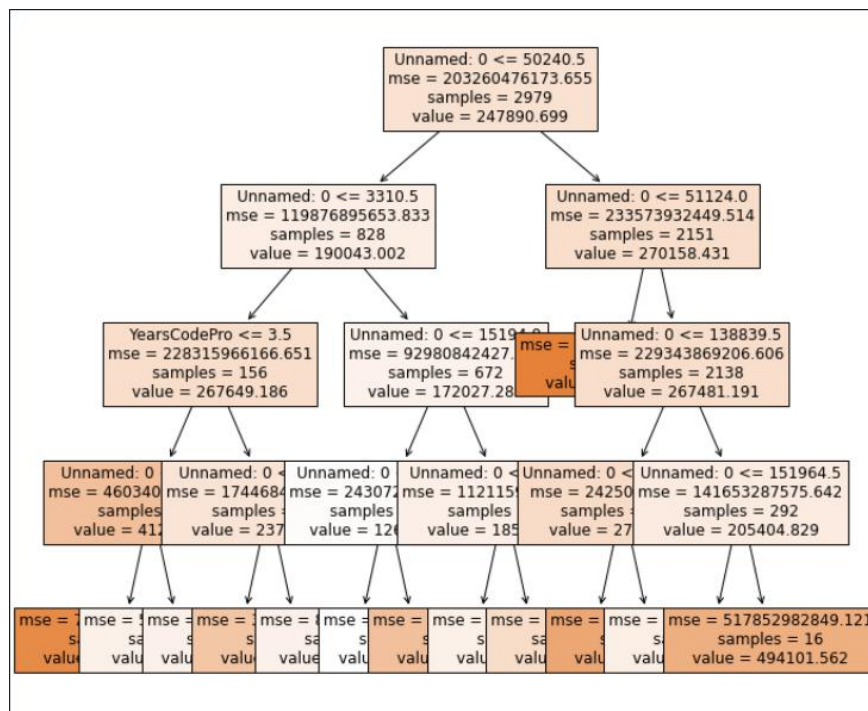
Our project splits the dataset into a training and testing dataset to validate the model. Generated randomly by machine, eighty percent of the data are used to test our model. Four models can be applied for this regression model: multiple linear regression, decision tree, KNN, and random forest.

The first model we made is multiple linear regression. We use all features to get an r-square is 0.234. It means that 23.4% of the data fit the regression model. Then, use the test dataset to validate the accuracy of the model. We use the MSE as the standard to evaluate the model. For linear regression, the MSE value is pretty high.

[illegible]

The second model we made is the decision tree. We put a range from 0 to 6300 with the sequence 700 to get the best impurity value to optimize the model. The best score from this range is -0.25439, while the minimum impurity value is 3500. For maximum depth, we've tried the range from 1 to 6. When the max depth hits 6, the MSE is relatively low.

```
min_impurity_decrease=0.000000 score=-1.383731
min_impurity_decrease=700.000000 score=-1.373863
min_impurity_decrease=1400.000000 score=-1.340666
min_impurity_decrease=2100.000000 score=-1.328976
min_impurity_decrease=2800.000000 score=-1.274472
min_impurity_decrease=3500.000000 score=-1.254390
min_impurity_decrease=4200.000000 score=-1.307897
min_impurity_decrease=4900.000000 score=-1.267659
min_impurity_decrease=5600.000000 score=-1.374519
min_impurity_decrease=6300.000000 score=-1.270741
```

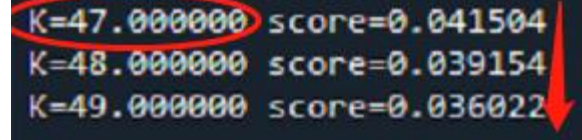


The third model we made is K-Nearest Neighbors. This algorithm is to calculate the similarity by distance. What we need to do is to set an appropriate parameter. To find the best K-value, I use the stimulation ranging from 1 to 100 with one sequence. The results show that there is one value. When $k = 47$, the score hits the peak. I test this number and find that the model has a smaller MSE when $k = 47$. And score begins to decrease when the k-value is more significant than 47. Thus, it is reasonable to assume that this parameter is most suitable for the model.

```

K=37.000000 score=0.032890
K=38.000000 score=0.036805
K=39.000000 score=0.030540
K=40.000000 score=0.033673
K=41.000000 score=0.032106
K=42.000000 score=0.030540
K=43.000000 score=0.036022
K=44.000000 score=0.036022
K=45.000000 score=0.032106
K=46.000000 score=0.034456
K=47.000000 score=0.041504
K=48.000000 score=0.039154
K=49.000000 score=0.036022

```



The last one we made is Random Forest. The random forest is a collection of unpruned decision trees, which can offset some outliers and noise. To get the best parameter for the random forest, we use the loop function to test the model's score from 1 to 30. The result shows that the max depth for the random forest is 1. As the picture listed above, it is easy to find that score decreases when the value hits 2. Thus, parameter one should be the best value for this model.

```

max_depth_increase=1.000000 score=0.002916
max_depth_increase=2.000000 score=-0.000239
max_depth_increase=3.000000 score=0.000518
max_depth_increase=4.000000 score=-0.002019
max_depth_increase=5.000000 score=-0.002610
max_depth_increase=6.000000 score=-0.009717
max_depth_increase=7.000000 score=-0.015074
max_depth_increase=8.000000 score=-0.016731
max_depth_increase=9.000000 score=-0.039703
max_depth_increase=10.000000 score=-0.043527
max_depth_increase=11.000000 score=-0.054390
max_depth_increase=12.000000 score=-0.069712
max_depth_increase=13.000000 score=-0.088921
max_depth_increase=14.000000 score=-0.098523
max_depth_increase=15.000000 score=-0.117476
max_depth_increase=16.000000 score=-0.121748
max_depth_increase=17.000000 score=-0.134656
max_depth_increase=18.000000 score=-0.148847
max_depth_increase=19.000000 score=-0.163122
max_depth_increase=20.000000 score=-0.165608
max_depth_increase=21.000000 score=-0.183668
max_depth_increase=22.000000 score=-0.185428
max_depth_increase=23.000000 score=-0.195962
max_depth_increase=24.000000 score=-0.187441
max_depth_increase=25.000000 score=-0.195056
max_depth_increase=26.000000 score=-0.201752
max_depth_increase=27.000000 score=-0.184292
max_depth_increase=28.000000 score=-0.186400
max_depth_increase=29.000000 score=-0.211691

```


Evaluation

<i>Model</i>	<i>MSE</i>	<i>MAE</i>	<i>R-Squared</i>	<i>RMSE</i>
Linear Regression	1.7149e+09	31637.5	0.239	41,411.351
Decision tree	1.835327e+11	234256.701401	-0.014704	428,407.166
K-NN	8.856576e+11	510349.726703	-3.896568	941,093.832
Random Forest	1.839079e+11	236494.233645	-0.016778	428,844.844

According to the four models we have selected, the linear regression model is the most suitable model according to the results of our current data. The smaller the value of MAE, the better the model. At the same time, the larger the value of R-square, the more significant the proportion of the data set that the model can explain. But R-square less than zero indicates that the model may be inferior to the baseline model.

Business Insights

Using this Linear Regression, the company's compensation committee can reevaluate the current data-related employee and adopt an appropriate salary guideline for the new interviewers. Committee officers will get the unique prediction based on the applicant's background by inputting certain employee information. On the other way, this model effectively avoids the waste of salary and increases employees' motivation. The employee is willing to learn more knowledge and devote themselves to the company because the company gives them positive feedback on salary.

Analyze Flaws

The model we made this time is not perfect. After our analysis, we concluded that there might be the following problems. First, our sample size is too small. Since we filtered out the observations in the United States and the salary base was more significant than 13,920 from the

data set, there were only about 5,000 observations left in the final data. In addition, it is also possible that when we delete Null values, we delete some comments with valid information. Meanwhile, we also believe that the data in the modified dataset is inaccurate because some people may think that this involves personal privacy, so they do not fill in it carefully. However, we believe that the analytical method we have deployed is not problematic.