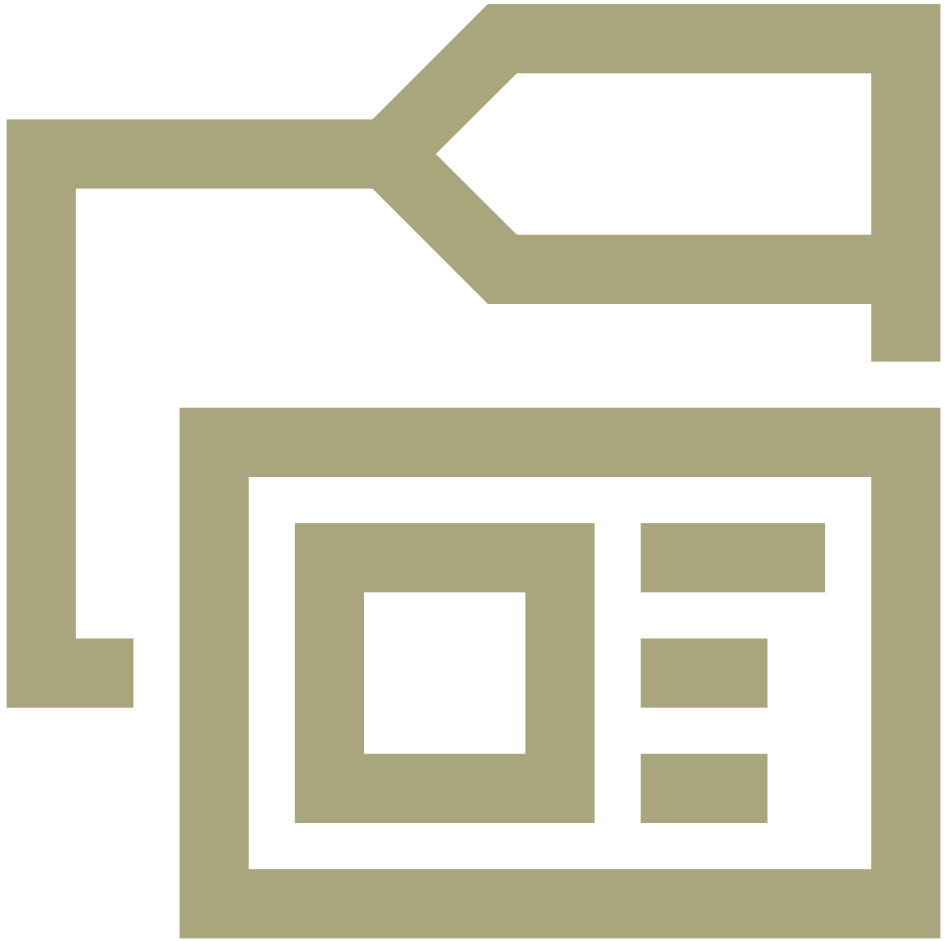
Data Scientists Salary Classification

Shreyas Choudhary



Data Cleaning

- The dataset contained many unnecessary features, especially in the other_text columns of the multiple choice responses. The values in that columns appeared to be -1 as they contained a free-flowing text and had been shuffled, so they were not in relation with anything. So these columns were dropped directly. Next, Q12 - Who/what are your favorite media sources that report on data science topics? And Q19- What programming language would you recommend an aspiring data scientist to learn first? were also dropped directly they do not contribute in any meaningful way to our salary classification analysis.
- Besides this, many columns contained null values, which could have been possibly resulted from either the respondents not filling the whole survey or they did not find anything relatable, for instance, if someone's education was only limited to high school, all other further questions related to educational qualifications became pointless to respond to.
- Further, instead of directly dropping the columns, they were replaced with mode, as it helped in reducing the loss of the information, which would be the case if the null values would have been dropped directly.

Exploratory Data Analysis

- Exploratory data analysis was done mainly on three parameters of country, age and educational qualifications. In order to be able to visualize the data, the whole data was grouped into suitable classes/categories and was then used. Some of the main findings from the exploratory data analysis were as follows:

i. The number of respondents of the survey majorly belonged to India and United States of America, were of the age group between 25-29 and had the highest educational qualifications of at least Master's degree.

ii. The continent of North America had the highest salary brackets among all the continents. Also, the trend of higher salaries with an increase in age group was seen, with age group 60-69, and educational qualifications of Master's degree and PhDs having highest salaries. (Exception was age group 70+, which is justified as it indicates retirement age.)

Feature selection and Feature Importance

- Random Forest Algorithm was used for feature selection, because by averaging out the impact of several decision trees, random forests tend to improve prediction. This ,on implementation reduced the number of columns from 319 to 140. For feature importance , mutual info from the scikit library was used. Mutual information (MI) between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency. Some of the highlighted important features included Q3_United States of America, Q3_India, Q9_Part 3, Q11_>100,000 USD as they had more scores as compared to other features on the feature importance plot.
- Further, after RandomForestClassifier, correlation with the target feature was checked and the threshold for the correlation was kept above 0.2 as anything below it will indicate that the feature is not directly correlated with the target variable. This filter gave us features Q3,Q8,Q9,Q11,Q15,Q23 as the most important ones with respect to the target variables. These were included with all their columns and thus a final feature engineered dataset was formed, which was going to be used further for the model implementation.

Model Implementation

Since the target for the dataset contained labels from 0-14 ordered in the increasing number importance hence implementing ordinary logistic regression model here wasn't feasible. Instead, an ordinal logistic regression model was implemented 14 times, each time with labels falling in different classes (0&1).

The accuracy of the model was calculated to be 36.84%. For calculating the accuracy , probability prediction of each model implemented in the 14 loops was stored in a list and compared with the test/validation set for the target variable.

Hyperparameter Tuning

Hyperparameters C and solver method were tuned for each model implemented. It was observed that there was trend of increase in the accuracy as we proceeded with the models, with the least of 77.3% accuracy for model_0 and 98.224% accuracy for model_14 in hyperparameter tuning.

The hyperparameters obtained from tuning were then implemented for each model accordingly and the overall ordinal logistic regression model accuracy score was then calculated again. This time the accuracy score went down to 30.96%.

Discussion(1)

- The decrease in the accuracy score was seen after implementing the hyperparameter tuning and 10-fold cross validation.
- The only way the decrease in the accuracy after hyperparameter tuning and k fold cross validation can be justified is by the fact that the data is overfitting.
- We have implemented the logistic regression 14 times and have used k fold cross validation with k=10 (split of data- 80% training and 20% testing),repeated the implementation 10 times with 10 different test data splits (each time a new 20% of all data).Therefore the result (in the case of hyperparameter tuning and cross validation) is more accurate and somewhat different than just one time split (as done in the model implementation) and may therefore account for the decrease in accuracy.
- The accuracy of such a model in general can be low because the model is repeatedly tested against unseen data all the time, and the final accuracy score is the mean of the individual test scores.

Discussion(2)

Accuracy can be increased by applying one or more of the following:

Exploring more classifiers - Logistic Regression learns from a linear decision surface that separates the classes. It may be the case that 14 classes given here may not be separated linearly. In such a case we might need to take a look at other classifiers such as Support Vector Machines which can deal with more complex decision boundaries.

Error Analysis - We might find that some of the models work well with a set of parameters while the other don't. In this case, Ensemble Techniques (such as VotingClassifier) often give the best results.

More Features – More features can also be included in the dataset in order to account for an increase in the accuracy.

Discussion(3)

The Bias- Variance tradeoff in this case can be given from taking a look at the confusion matrix and the classification report of the model implemented. Logistic Regression model is a high-bias, low-variance model in general and it can be seen that it is having high precision and low recall from the classification report.

A model might not end up doing well on the training data but it converges better. Such a model would have a higher bias and lower variance, which is our case.

We can have a model which gets some false negatives but gets fewer false positives, i.e., it is high precision - low recall, thus corresponding to the high bias - low variance case.