This assignment is aimed to train, validate, and tune multi-class ordinary classification models which can classify what a survey respondent's current yearly compensation bucket is by given a set of survey responses by one data scientist.

1. Data Cleaning

We are going to have the encode the whole data frame to get a clean data frame. First, I dropped the first column "Time from Start to Finish (seconds)" which has no relation with our target current yearly compensation, and also column "Q25" and "Q25_buckets" since we already have the encode column "Q25_Encoded" for later analysis. While checking the survey result, I found there are many answers such as "Other" are too blurry for our study. So I change all "Other" value in the data frame to missing value NaN, which I will encode it later. I also changed the respond "Pefer not to say" and "Prefer to self-describe" to NaN in the gender column. Then, since all parts related with Q27B, Q29B, Q30B, Q31B, Q32B, Q34B, Q36B, Q37B and Q38B are asking about features about future two years, which seems to have very weak relation with our target current yearly compensation, so I dropped these columns.

Some questions in the survey are select-all-apply questions, so I decided to add all parts for one question together to see the actual number of responds. I created a new data frame to store the sum of these type of questions, and then get rid of the questions which have too many missing values. The threshold I set here is 8000, which is nearly the half of the sample. So now, Q18, Q19, Q29A, Q30A, Q32A and Q37A are removed since there were over 8000 missing values. The possible reason may be Question 18 and 19 were only asked to respondents that selected the relevant answer choices for Question 17. Question 29-A and 30-A were only asked to respondents that selected more than one choice for Question 27-A. After that, we are going to find the total missing value for the rest questions. The threshold is still 8000, and finally Q1, Q2, Q3, Q4, Q5, Q6, Q8, Q11, Q13, Q15, Q20, Q21, Q22, Q23, Q26 and Q41were kept. We can see in all multiple choices with one answer, only "Q28"(Of the cloud platforms that you are familiar with, which has the best developer experience (most enjoyable to use)), "Q33" (Which of the following big data products (relational database, data warehouse, data lake, or similar) do you use most often?) and "Q35" (Which of the following business intelligence tools do you use most often?) has more than 8000 missing values, so we will drop this column. The possible reason is Question 28, 33 and 35 was only asked to respondents that selected more than one choice for Question 27-A, 32-A and 34-A respectively where Question 32-A already had too many missing values.

Then, I changed categorical variables to numerical which applies ordinal encoding for responds with more than 2 levels, and also apply one-hot encoding on select-all-apply questions. Here, I used the factorize function to do the ordinal encode, which cannot obtain the certain order of features, such as 'Never' as 0, 'Once' as 1 and '2-5 times' as 2 and so on, which may be little bit confusion in the coefficient plot part.

2. Exploratory Data Analysis and Feature Selection

To visualize the order of feature importance, I used linear regression model. Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables, where current yearly compensation is the scalar response and other questions are determined as explanatory variables. Based on the feature importance plot (Figure1), I plot the coefficient of the model with target variable is current yearly compensation. We can see the variance is between coefficient is large, and not all coefficients have large enough impact on the target. So, the threshold interval is set at [-0.1, 0.1] to avoid low coefficients that has minor influence on the target value. Since there are too much missing value in some parts of select-all questions taken into account, which will cause large variance, so we will select these features for the further analysis: Q2, Q4, Q5, Q6, Q11, Q21, Q22, Q23, Q26, Q7_Part_1, Q14_Part_1, Q14_Part_2, Q16_Part_1, Q24_Part_1 and Q42_Part_4.

3. Model Implementation

In this question, I implemented ordinal logistic regression algorithm on the training data by using 10-fold cross-validation. Ordinal logistic regression is a statistical analysis method that can be used to model the relationship between an ordinal response variable and one or more explanatory variables. Repeated k-fold cross-validation has the benefit of improving the estimate of the mean model performance. The disadvantage of k-fold cross-validation is that it can be slow to execute, and it can be hard to parallelize. First, I defined all selected features as X and Q25_Encoded as y. Then, I split the dataset into 70% training set and 30% test set. Normalization is applied here, since it can help to put all parameters on the same scale, so that some are not massive while others are tiny. This is something that can cause trouble with fitting models. Also, it frequently helps with interpretation of and comparison between each of the estimated parameters. The accuracy of this model is 83.01% on the training set by using the formula $Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$. By using 10-fold cross-validation, we can get 10 probabilities of the output belonging to each of the salary buckets, which are 0.82, 0.82, 0.806, 0.831, 0.823, 0.855, 0.811, 0.815, 0.814 and 0.827 respectively. As a result, the average accuracy is 82.23% with variance at 1.29.

4. Model Tuning

To find the model with optimal performance, here I applied grid search method. Grid search is a tuning technique that attempts to compute the optimum values of hyperparameters. Value of C is set to 0.001,0.01,0.1,1,10, 100 respectively while solver is set to “newton-cg”, “lbfgs”, “liblinear”, “sag” respectively. The result shows the optimal logistic model uses C = 0.001 with a liblinear solver and has a cross validation score of 82.26% with a standard deviation of 1.27. Confusion matrix is used here, since it is a very intuitive cross tab of actual class values and predicted class values. It contains the count of observations that fall in each category. Accuracy score is the ratio of correct predictions to total predictions. I plot the feature importance plot by using the correlation between target value and features on the training set. From the following graph, we can see Q6, Q11, Q22, Q23, Q26, Q24_Part_1 have outstanding influence on target value. Compare with figure1, less features are having outstanding influence on current yearly compensation.

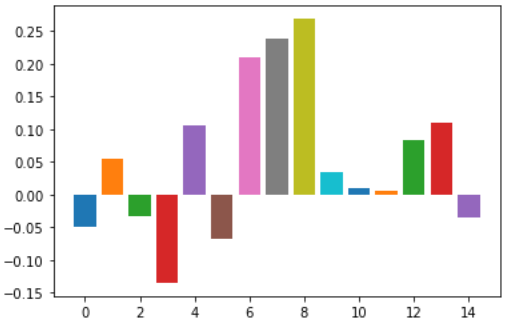


Figure2. feature importance plot (Q4)

5. Testing and Discussion

The optimal model with C = 0.001 with a liblinear solver has accuracy at 82.26% with standard deviation at 1.27. I think the model is acceptable. Based on Figure3 and Figure4, we can find that the model is not overfitting since the training set is not performing better than the test set, where overfitting is a modeling error that introduces bias to the model because it is too closely related to the data set. A statistical model is said to have underfitting when it cannot capture the underlying trend of the data. Underfitting destroys the accuracy of model. Its occurrence simply means that our model or the algorithm does not fit the data well enough. It usually happens when we have fewer data to build an accurate model and when we try to build a linear model with fewer non-linear data. Here, we may increase the size of training set and also reducing the features by feature selection to avoid underfitting and can increase the accuracy as well. Cross validation is also important to improve the accuracy of the model. From the data set and trained classification model, I learned that gender, level of education, current role, programming experience, using python on regular basis, type of computing platform regularly used, using matplotlib or seaborn on regular basis, using Scikit-learn machine learning framework on regular basis, size of company, individuals are responsible for data science at working place, current employer incorporate machine learning method, analyze and understand data to influence product or business decisions ability, money spent on machine learning services and use Kaggle as media sources on data science topic are important features that influence annual salary.

Appendix

Figure1. feature importance plot (Q2)

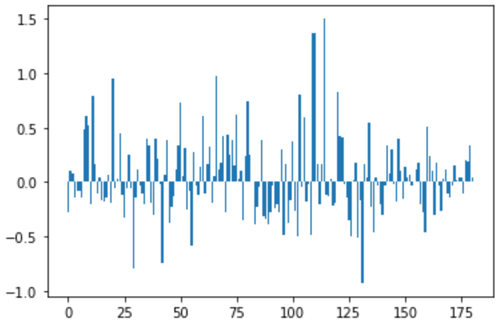


Figure3. target variable values VS predictions on training set (Q5)

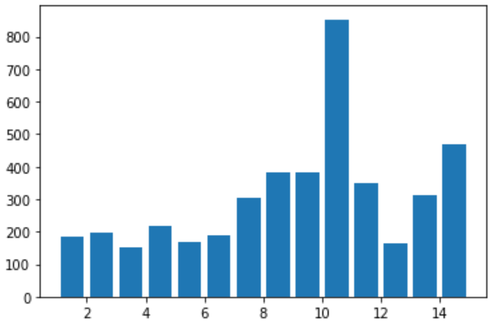
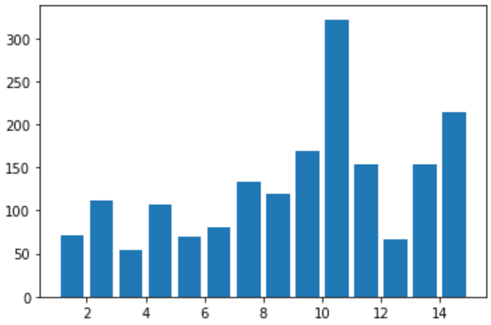


Figure4. target variable values VS predictions on test set (Q5)


300
250
200
150
100
50
0
2
4
6
8
10
12
14