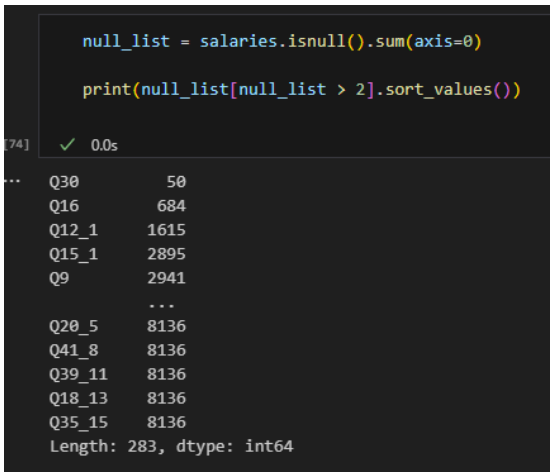
1. Data cleaning:

1.1 Drop Data:



After a brief inspection of the dataset, I decided to drop column ‘Q29” because it was already encoded in the ‘Q29_Encoded’ column. Next, I dropped the ‘Duration’ column since it’s not relevant to the task. In addition, I added an extra rule to perform a drop action at row level if a row contains more than 90% of null value.

1.2 Handling Missing Values:

First, I print out the total number of missing values in each column and sort the list in ascending order. According to the null list, I decided to simply drop the null value on columns that have less than 100 missing data. Then, fill the column with mode or mean for columns that have more than 100 and less than 1000 missing data. For other columns, I decided to replace the null value with ‘Unknown’ (for single response) and ‘Not Select’ (for multiple response).

1.3 Encoding Features:

After all the missing values have been imputed, I applied the label encode function on single response columns to encode categorical data into numerical values. Next, I replace the value in multiple response columns with 0 if the value equals ‘Not Select’ and 1 for other options. Then, I write an algorithm to automatically merge multiple response columns who share a common name before the ‘_’ underscore sign. As a result, I obtained a cleaned data frame with all numerical values and ready for the next step.

2.Exploratory data analysis and feature selection:

2.1 Feature Selection:

I use the Lasso regression model for feature selections because the lasso linear model tends to produce sparse models by dividing coefficients to exactly zero. The method is extremely useful when you trying select the relevant features from high-dimensional datasets. Most importantly, lasso models use a regularization term to penalize the absolute value of coefficients which present overfitting and lead to a more interpretable result.

3.Model Implementation:

3.1 Cross-validation:

After performing a 10-fold cross validation on my training data. I obtained a average accuracy of 34.954% with variance 0.0002824

3.2 Scaling/normalization of features:

Ordinal logistic regression is somewhat robust to the scale of predictors, as it estimates separate slopes for each category. Additionally, the dataset after encoding does not show a great difference between individual values. Therefore, the step of normalization is not necessary for this task.

4.Model tuning:

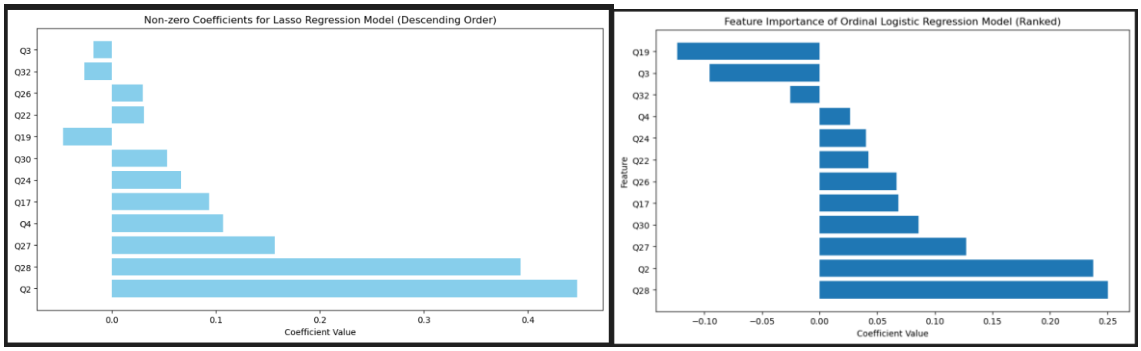
4.1 Evaluation Metric:

In an ordinal logistic regression model, the results tend to present with an inherent order or ranking. However, accuracy is a metric that treats all misclassifications equally and the effect of ranking becomes pointless. Furthermore, the class has different importance in ordinal logistic regression which accuracy takes account when degree of different between classes.

4.2 Hyperparameter Tuning:

I selected max_iter and C as my hyperparameter. After applying grid search technique, I obtained the best set of hyperparameter to be {'C': 0.001, 'max_iter': 100} with a accuracy of 35.39%.

4.3 Feature Importance Comparison:



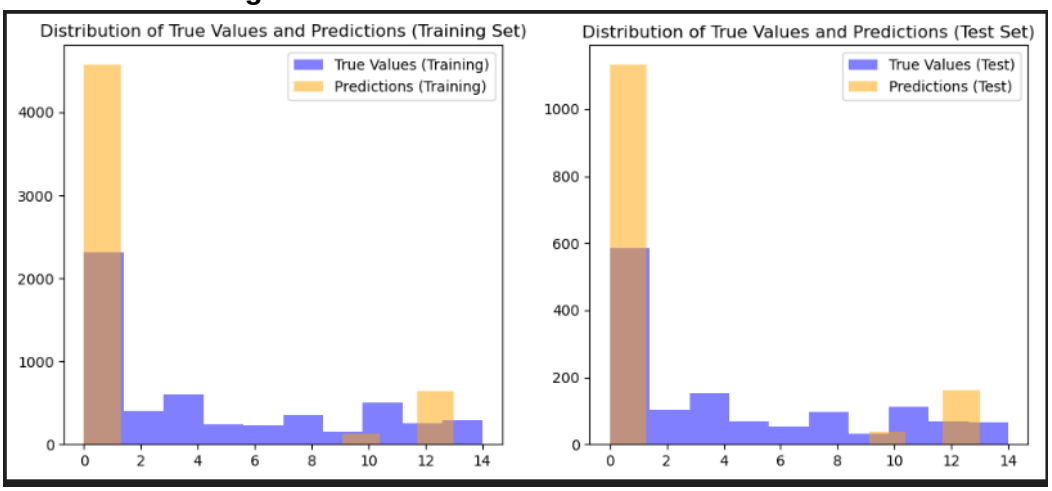
In comparing Lasso Regression and Ordinal Logistic Regression with Recursive Feature Elimination (RFE) for feature importance, Lasso emphasizes individual coefficients' magnitudes, driving some to zero for inherent feature selection. The graph displays coefficients in descending order. On the country, RFE in Ordinal Logistic Regression ranks and selects features based on their cumulative impact on model performance. The graph shows features ranked by importance. Lasso emphasizes individual features, while RFE focuses on overall contributions. The choice depends on interpretability needs and goals for the specific dataset and model.

5.Testing & Discussion:

5.1 Model performance

The accuracy of your best-performing model is approximately 35.39% on the training set and 36% on the test set. These results suggest room for improvement, and it's recommended to explore adjustments such as hyperparameter tuning or trying different algorithms to enhance model performance. Consider evaluating additional metrics like precision, recall, or F1 score for a more comprehensive understanding of your model's capabilities.

5.2 Discussion on fitting results:



According to the resulting diagram, a clear underfitting can be observed. To address this problem, I have a few methods to attempt. First, I can use another encoding technique which will include all responses of a single multiple-response column. For example, I merge the response for a same question into a column that contains the number of valid responses. On another hand, I can display the merge column as a vector [1,0,1,0,1] where 1 represents selected and 0 as not select. By applying this method, the model should find a more suitable pattern for each muti-response column. Secondly, I can add more hyperparameters into the model and perform hyperparameter tuning at a bigger scale. With more hyperparameter added to the model, the model should be able to adapt into dataset with more flexibility.

5.3 Insights and Discussion:

In examining the extensive survey response dataset, crucial insights have been discovered. Particularly, features wielding the most influence in forecasting a respondent's annual compensation were discerned through a ordinal logistic regression model and meticulous model tuning. The impact of skills, experience, and specific survey responses on compensation levels became evident. Rigorous evaluations on both the training and test sets shed light on the best-performing model's generalization capabilities. Dive into the intricacies of the bias-variance trade-off through hyperparameter adjustments provided nuanced insights for optimizing model performance. Additionally, a thorough exploration of the feature importance comparison, unraveling the pros and cons of ordinal logistic regression model. While the analysis unearthed valuable insights, it's essential to acknowledge inherent limitations of the process method to dataset and the chosen approach.. In conclusion, the implications of these insights extend beyond the analysis itself, offering practical guidance for survey dataset transformation and logistic regression implementation.