Part 1

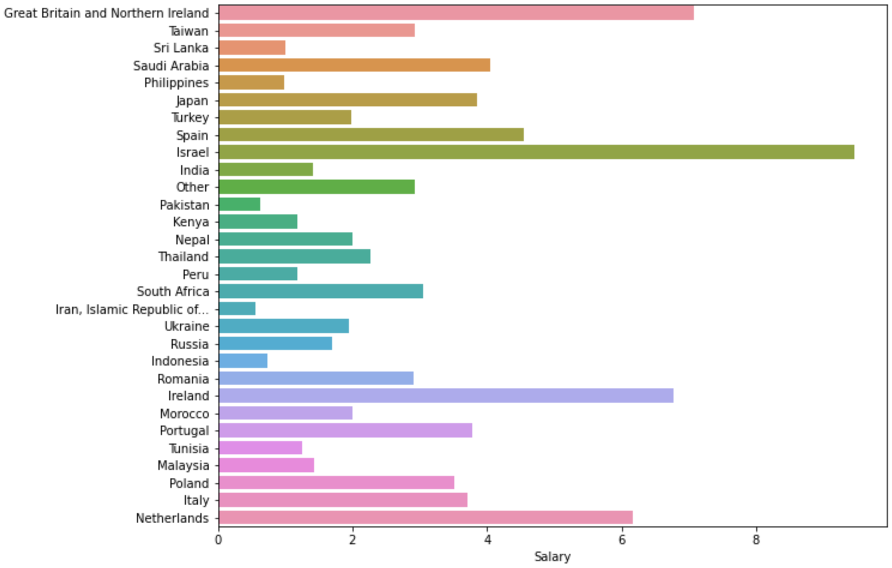
Data Cleaning and Encoding

(i) The questions which asked 'Select all that apply' i.e. Multiple Choice Questions (MCQs) were treated differently than the rest. This is because a respondent may or may not select a particular option which may reflect as null in dataset. This does not mean the value is missing, however, it just means it was not selected.
For such questions, their options were treated as separate independent features which may have values 1 implying selected, and 0 implying not selected.

(ii) Among Non-MCQ questions, questions Q22, Q32 and Q43 had more than 50% null values, so they were removed, as this is very less availability to impute these features without bias. Q5 had only one unique response which makes it incapable to act as a predictor, hence was removed. Q29 which is target variable in slightly different form, was removed as targets should never be in training data in any form.

(iii) For Q9, Q16 and Q30, the null values were replaced with most frequent responses as this seem reasonable in the context of these questions.

(iv) As can be seen below for a subset of countries, there was a clear dependence of average salary on the country of residence.



However, the number of countries were too large to encode them separately. So, they were grouped into four major categories with label encoding whose numerical values were higher for countries with tendency of higher salaries.
For example, USA was encoded 4, Great Britain 3, Spain 2 and Kenya 1

(v) For all the Non-MCQ questions, except 'Current Role' and 'Industry', the categories were encoded in such a way that their numerical values represented an inherent order in their context.

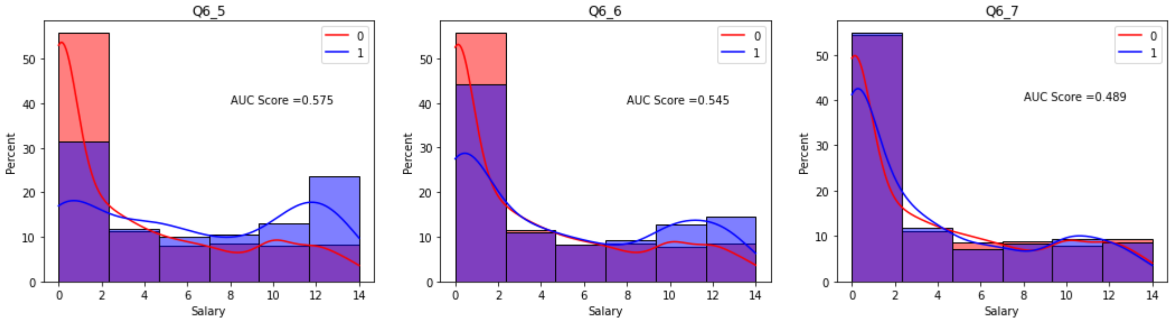
Part 2

Exploratory Data Analysis and Feature Selection

(i) Feature selection and dimensionality reduction should be independent of test data. Thus, it was appropriate to split the dataset into training and test set at this stage.

(ii) The training set was then split into MCQ and Non-MCQ questions as they had to be handled separately as is discussed in next steps.

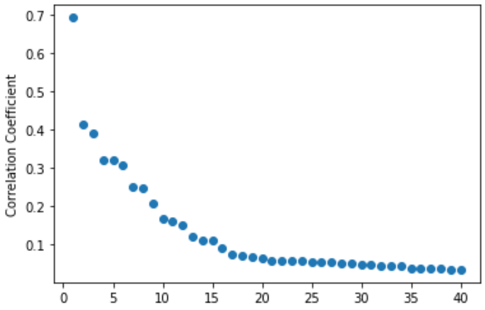
(iii) For MCQ features which have values 0 or 1, we performed feature selection using AUC score. This technique basically gives a quantitative estimate of the difference in salary distribution corresponding to value 0 and 1. This is illustrated in below figure,



Using this technique, 126 MCQ features were selected out of 277 in total.

(iv) Next, among the selected non-MCQ and MCQ features, dimensionality reduction using MCA technique was performed on those features which were purely categorical. These include all the MCQ features, gender, current role, and industry from the non-MCQ features.
This separation for MCA was done because the MCA algorithm works by transforming all categories into one hot encoding, which in principle means numerical values will lose their significance as they will be treated only as categories. MCA helped in dimensionality reduction from 129 such categorical features to 80 features which explained 75% of variance. Refer Appendix A.

(v) Finally, feature importance analysis was performed on above 80 (after MCA) + 10 (from non-MCQ not included for MCA) i.e., 90 features. Of these 90 features, 40 features were selected for use in the classification model based on following correlation analysis with target variable,

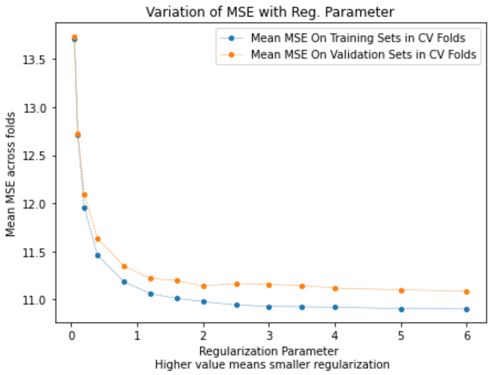


Part 3

Model Implementation

(i) Ordinal logistic regression algorithm was implemented by performing multiple one vs rest logistic regressions and individual probabilities were obtained by successive subtraction of probabilities.

(ii) 10-fold cross-validation was done by changing just one hyperparameter for regularization,



Part 4

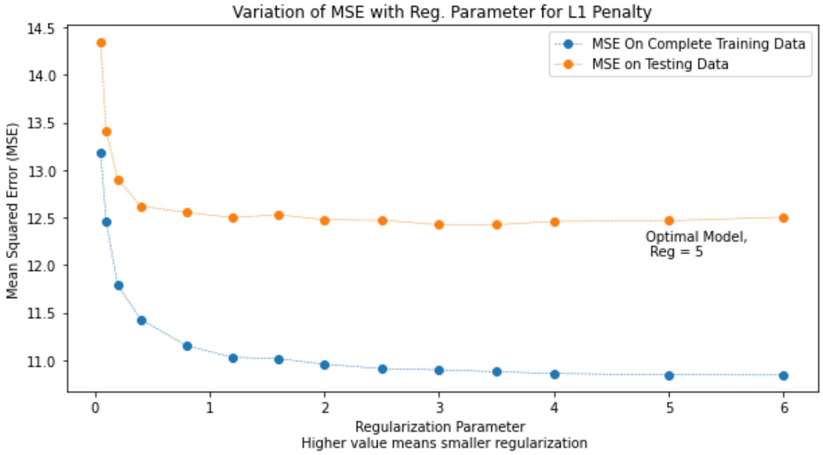
Model Tuning

(i) Hyperparameter tuning for penalty type and reg. parameter was done using grid search and best model was found based on lowest mean MSE on validation sets during cross validation, which was penalty type : L1, and Regularization Parameter value : 5.
Please see Appendix A for change in order of feature importance before and after model tuning.

Part 5

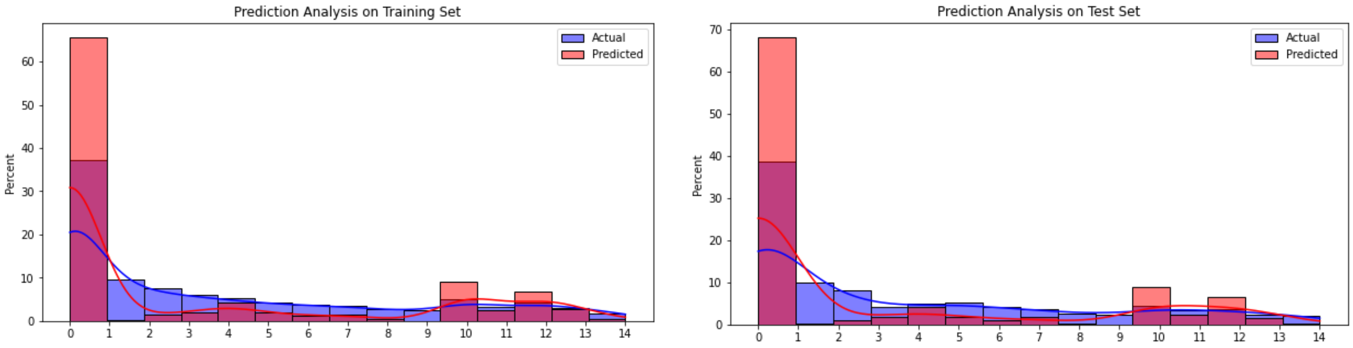
Testing and Discussion

(i) Classification performance is compared on complete training data and test data (unseen by model until now) for the best model, and for models with different regularization for reference as below,



Above plot shows that the performance of optimal model is adequate on test data as well. Also, this model seems to be a good fit because the test error is not increasing at this value, and testing and training error are stabilized to a minimum.

(ii) The actual v predicted distributions of salary classes for training and test set are as follows:

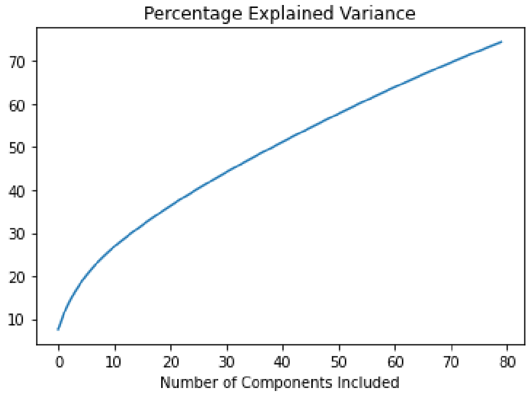


We observe that the prediction performance of the optimal model is very similar for both training and test data. This indicates we have neither overfit, nor underfit.
Secondly, though the predicted classes are close to actual classes, but the major difference can be seen for the salary buckets -0,1,2. Thus, major cause of low accuracy of the model is predicting salary class 0 instead of 1 or 2. However, considering the fact that these classes are closer to each other , it is safe to say that the model does a satisfactory job in predicting approximate salary classes i.e., salary classes close to the true salary classes. This performance reflects the effectiveness of ordinal logistic regression for target variables which possess an inherent order.

Appendix A

Part 2

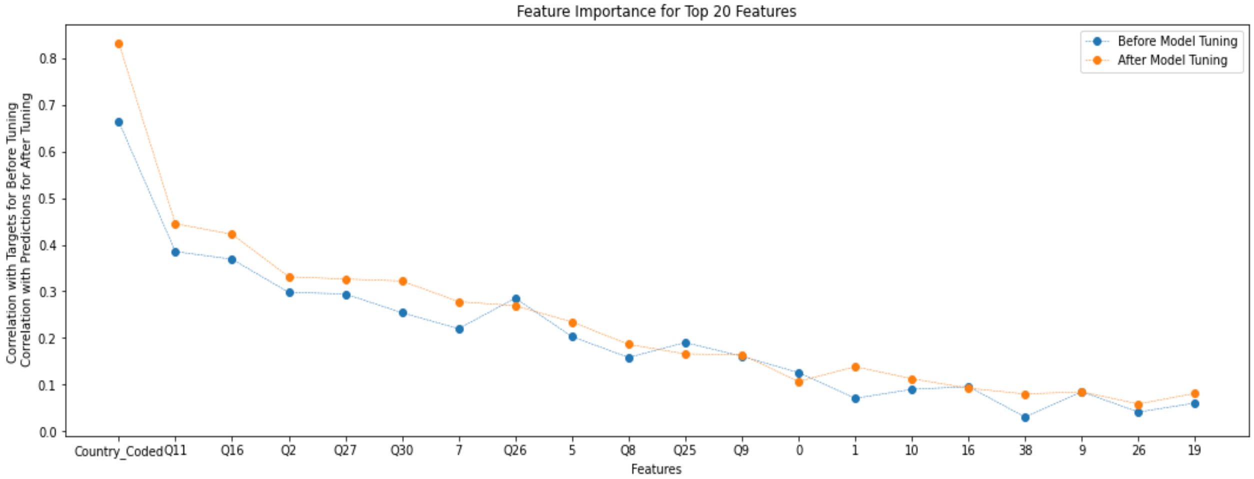
Cumulative explained variance from MCA for 129 categorical features when reduced to 80 features, is as below,



This shows the reduced dataset with 80 features out of 129 features have 75% of the original available information.

Part 4

Feature Importance before and after model tuning is as below,


Feature Importance for Top 20 Features
Before Model Tuning
After Model Tuning
Correlation with Targets for Before Tuning
Correlation with Predictions for After Tuning
Country_Coded
Q11
Q16
Q2
Q27
Q30
Q26
Q8
Q25
Q9
Features