ASSIGNMENT 2 REPORT-ORDINAL LOGISTIC REGRESSION

QUESTION 1: DATA CLEANING

These are the procedures taken in cleaning the dataset:

1. The first row of the original dataset was removed since it only describes each column; and the Q24 column was removed because it has been encoded into Q24_Encoded and Q24_buckets columns.
2. The single choice column was checked for the number of missing values. Columns Q38, Q30, and Q32 had more than 1000 missing values so they were dropped because too much information was missing.
3. Column Q8 was dropped as it shows the programming language each participant recommends for an aspiring data scientist. Intuitively, this column has nothing to do with salary as each participant recommendation can be based on factors such as, difficulty of the language.
4. All the parts for each multi-choice feature (say Q7_Part_1 to Q7_OTHER) was combined to determine how many participants fail to select at least one option.
5. The combined missing data for each multi-choice column was computed, the multi-choice features with more than 1000 combined missing data was removed because too much information was missing.
6. The entries of the 561 participants who failed to select an option for column Q11 were dropped because they failed to also select an option for the multi-choice column.
7. The remaining entries select at least one option in the multiple-choice columns, so no imputations were needed.
8. The missing values in Q25 was imputed with the mode of amount spent on machine learning services by yearly income of the participants because high salary earners were found to spend more on machine learning services.
9. The multiple-choice columns was converted to numerical data by assigning 1 to non-null values and 0 to null values.
10. Columns Q7_Part_12, Q9_Part_11, Q10_Part_13, Q12_Part_3, Q14_Part_11, and Q23_Part_7 was removed because they represent entries that selected 'None'.
11. Dummy encoding was used for nominal categorical columns to prevent multicollinearity between features and label encoding was used ordinal categorical columns to maintain the ranks between values.
12. The dimensions of some categorical data were reduced by selecting the top frequent categories then grouping other categories to others.

QUESTION 2: EXPLORATORY DATA ANALYSIS AND FEATURE SELECTION

Feature Engineering: summing up the multi-choice columns Q7, Q9, Q10, Q12, Q14, Q23 and their different parts depict how experienced each participant is in the field of Machine Learning and Data Science tools.
The distribution of the salary bucket in the cleaned dataset is unbalanced according to figure 1.

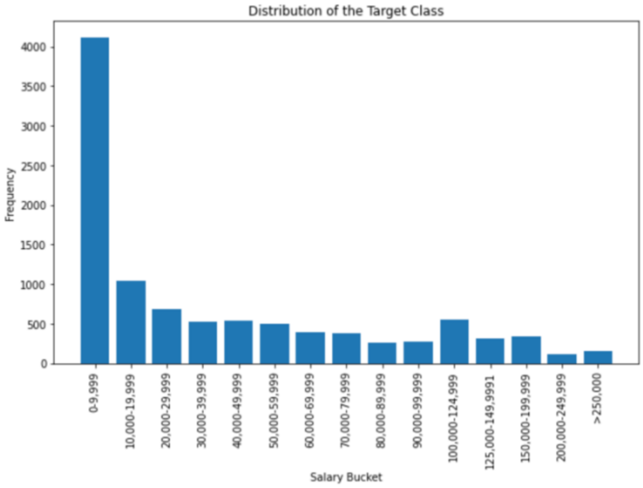

Figure 1: Distribution of the Target Class in the dataset.

Scaling of the predictors: scaling of the predictors is not necessary because all the predictors are categorical feature which are either label encoded or dummy encoded.

Splitting of the dataset was done before feature selection to avoid information leakage to the model.

Feature Selection Technique: Chi-Squared test was used for feature selection because it checks for the dependency of feature column on the target column. Independent features are not useful for prediction while the dependent features have predictive power on the target.

Feature Importance: features with lower p-values returned from the chi-square test are the significant features. Also, higher the chi-square score means a feature is more dependent on the target variable; hence can be selected for model training.

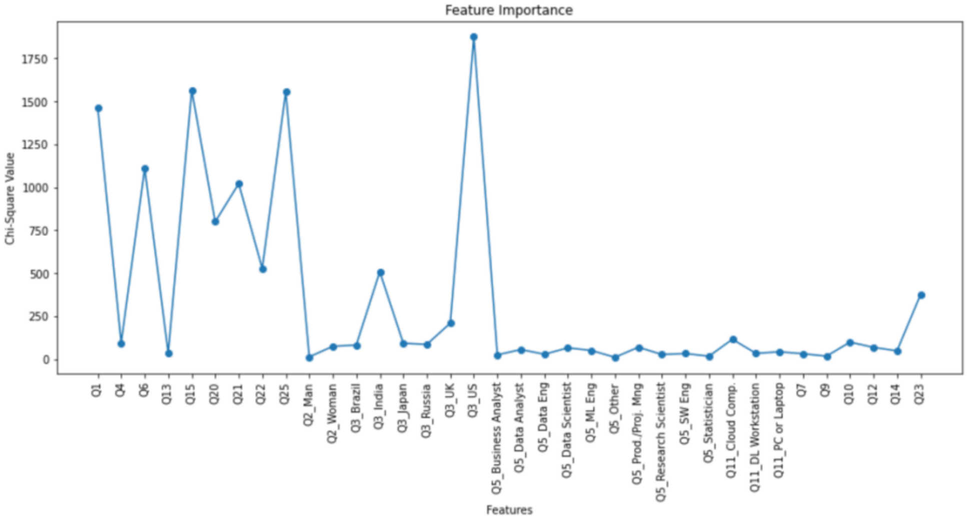

Figure 2: Feature Importance.

The most important feature according to the graph in figure 2 was Q3_US.

Selecting features: features were selected by setting alpha<=0.05 for their p-values. Features whose p-values are between 0 and 0.05 are dependent on the target variable (useful predictors) while those whose p-values are above 0.05 are independent of the target variable (useless predictors).

QUESTION 3 & 4: MODEL IMPLEMENTATION & HYPERPARAMETER TUNING

Implementation and Tuning of the Ordinal Logistic Regression (OLR) algorithm on the training dataset using 10-fold cross-validation.

Parameters of the ordinal logistic regression model:

C (Inverse Regularization Strength)- controls the trade-off between allowing the model to increase it's complexity as much as it wants with trying to keep it simple. High 'C' values allow the model to increase its complexity (overfit) while low values tend to underfit the model.

Penalty- (L1, L2)- model penalized by 'L1' is called Lasso Regression while model penalized by 'L2' is called Ridge Regression. Ridge regression adds squared magnitude of coefficient as penalty term to the loss function while Lasso regression adds absolute value of the magnitude of coefficients.

max_iter- maximum number of iterations taken for the solvers of the logistic regression to converge

Other Hyperparameters- solver, multi_class and so on.

Tuned Hyperparameters.

- C (Inverse Regularization Strength) – the 'C' parameter was tuned because to reduce the chance of overfitting or underfitting of the models.
- Penalty – 'L1' and 'L2' penalties adjust the coefficient of each parameters.

Performance Metrics (Accuracy): accuracy was used as performance metrics because the interestof the project is about the correct predictions (yearly compensation) the model can make, and not how wrongly it classifies other classes for cases like anomaly detection.

The optimal model after hyperparameter tuning had cross validation score of 43.586% with a standard deviation of 2.501% across the 10 folds at C=1, with 'L1' penalty.

The accuracy of the optimal model across the 10 folds ranges between 39.7% to 46.8%. The average accuracy of the optimal model was almost between the the minium and maximum accuracy achieved by the model across the 10 folds ((39.7)/(46.8)=43.25%); this indicates low-bias. The standard deviation of the accuracies across folds is low (2.501%) and this indicates low-variance. Figure 3 shows the probability distribution for a single entry across all salary buckets.

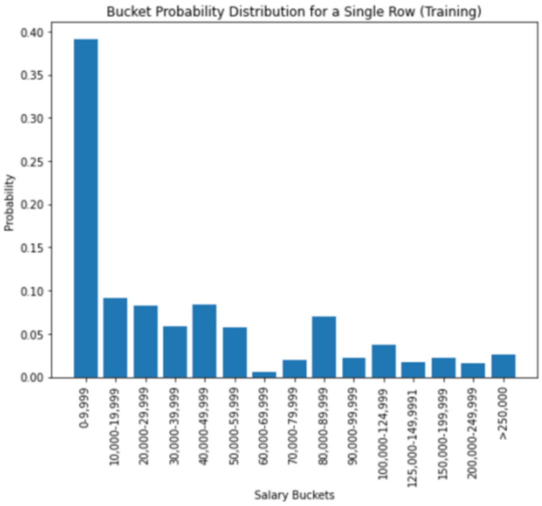

Figure 3: Probability distribution of bucket salary for a single entry.

QUESTION 5: TESTING AND DISCUSSION

The accuracy of the optimal model on the training set is: **44.44%** as compared to **41.92%** on the test set. The accuracy on the training set is just **2.52%** above the test set; this indicates that the model is not overfitting. The model is clearly underfitting as the model has low accuracy on both training and test dataset. This means the model can't capture the relationship between the explanatory data and the reponse data. The poor performance can be connected to the issue of unbalaced dataset in question 2. The performance of the model can be improved by adding new domain-specific features and getting more samples of other classes to balance the dataset.

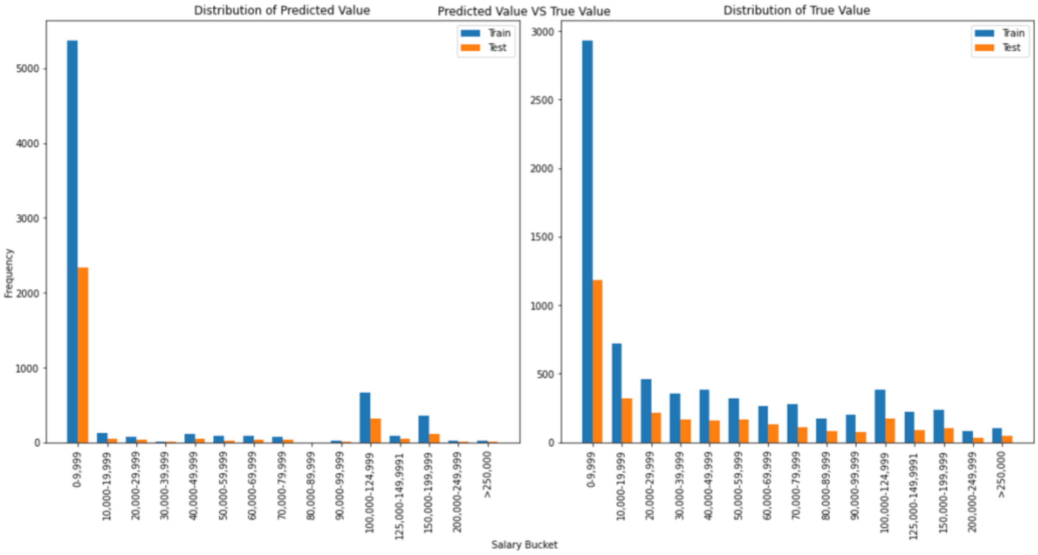

Figure 4: Distribution of Predicted Value VS True Value for Training and Test Set.

The dataset set used in the creation of the ordinal logistic regression model is unbalanced and this has a huge effect on the performance of the model. Figure 4 shows that the distribution plot for the predictions on the training and test set are similar which shows that the model generalizes on the features available to it as best as possible; providing more domain-specifc features to it will surely improve its performance.