MIE1624 Assignment 2 Report

Ziqingqing Ye, 1009167544

November 16, 2022

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

Data Cleaning

Data cleaning is completed through following steps:

- Drop dataframe columns based on NaN percentage. In this case, I dropped columns that contains 80% or more NaN values. Those columns don't contain a lot of information that helps analysis, dropping them will be fine. After
- Drop unnecessary features: 'Duration', 'Q29', 'Q29_buckets'. Since the response time is irrelevant in the survey, thus it can be removed. Also, Q24 and Q24_buckets are already encoded and represented by Q24_Encoded, so they can be removed as well.
- Drop the "Other" columns. For example, 'Q6_12'. There are choices "Other" in multiple choice questions. Other can be deleted since it doesn't contain useful information.
- Encode categorical feature with order. For the categorical feature that has order, I firstly replaced all the NaNs by mode values and then encoded them from low to high. Categorical features with orders 'Q8 What is the highest level of formal education', 'Q11 For how many years have you been writing code', 'Q43 Approximately how many times have you used a TPU', 'Q16 For how many years have you used machine learning', 'Q25 What is the size of the company', 'Q26 Approximately how many individuals are responsible for data science workloads at workplace' and 'Q30 Approximately how much money have you spent on machine learning'. After doing this, I replaced all NaN values by 0 and non- NaN values by 1 for each multiple-choice answer.
- Encode all the remaining categorical data: 'Q2', 'Q3', 'Q4', 'Q5', 'Q9', 'Q23', 'Q24', 'Q27', 'Q32'. I encode all the remaining features by assign numbers to each category. For example, man is zero while woman is one.
- Drop the row with questions: Since the first row contains the description of questions, it can be removed.

Exploratory data analysis and feature selection

The correlation graph is displayed as a heat map. The cmap parameter is coolwarm, means a cooler color means a stronger negative correlation, while a warmer color means a stronger positive correlation. As shown in the correlation plot, 'Q4 Country' with correlation value around 0.5 that most closely related to a respondent's annual compensation. The four most important characteristics are "Country Q4", "Years programming Q11", "Years used method Q16"and "Age Q2".

Feature selection is important because it prepares an input data set that is compatible and better suited to the machine learning algorithm and also improves the performance of machine learning models. This task requires to select features that are highly dependent on the response variable so that irrelevant features can be removed. I used the Chi-Square Independence Test to determine if there are any characteristics that are independent of the target variables. The null hypothesis for this test is that there is no relationship between the function and the target variable yearly compensation. Alternative hypothesis is that there is a relationship between them. A chi-square test is calculated between each characteristic and the target variable. By setting the confidence level to 0.05, the characteristic is wage dependent if the p-value is less than 0.05; otherwise, the feature is salary independent and should be removed. Consequently, 'Q9', 'Q13_5', 'Q13_11','Q15_2', 'Q35_1', 'Q44_4', 'Q7_5', 'Q42_8', 'Q12_1', 'Q7_2', 'Q6_7', 'Q15_1' are negligible and removed functions.

Model implementation

In ordinal logistic regression, the dependent variable is ordinal. Multiple binary classification with orders is required before performing logistic regression. Since there are 15 classes in the salary target variable, the binary classification is made by separating the classes from the low to high. Let Y be an ordinal result with 15 categories. Then the cumulative probability of Y is less than or equal to some category j = 0, ..., 14. The probabilities for each class are calculated from the new cumulative probabilities minus the last cumulative probabilities. For example, I calculated the probabilities for Class 0, Class 0 + Class 1, Class 0 + Class 1 + Class 2, etc. And then I get probabilities for class 0, class 1, class 2, etc. After calculating the probability of each class, the class with the maximum probability is selected for each observation to complete the multi-class predictions. Therefore, the convolution accuracy values are quite similar and the average accuracy value is 39.192% with a variance of 0.044%.

Bias of an estimator equals to the expected value of the estimator minus true value. I choose to use squared bias here to avoid negative bias value. I tested parameter C in list [0.001,0.01,0.1,1,10,100] and also plot the bias variance trade-off. As the C change from 0.0001,0.001,0.01,1,10,100, the bias gets lower while variance gets larger. In order to decrease bias, complexity should be increased.

Model tuning

For the logistic regression function, the parameters are as following: 'penalty', 'dual', 'tol', 'C', 'fit_intercept', 'intercept_scaling', 'class_weight', 'random_state', 'solver', 'max_iter', 'multi_class', 'verbose', 'warm_start', 'n_jobs', 'l1_ratio'.

In this assignment, I choose C and Solver for model tuning.

- C: The parameter C the inverse of regularization strength. For small values of C, we increase the regularization strength, resulting in simple models that do not fit the data well. For large values of C, we lower the regularization strength. The default value is 1.0. For tuning we will try C = [0.001,0.01,0.05,0.1,0.5,1,5,10, 100].
- Solver: Algorithm to use in the optimization problem. Algorithm includes [‘newton_cg’, ’lbfgs’, ‘liblinear’, ‘sag’, ‘saga’]. Default = ‘lbfgs’. For tuning we will try solver = ['newton cg','lbfgs','liblinear','sag'].

Grid search is the process of performing hyperparameter tuning in order to determine the optimal values for a given model. In this case, I pass predefined values for hyperparameters "C" and "Solver" to the GridSearchCV function. GridSearchCV tries all the combinations of the values passed in the dictionary and evaluates the model for each combination using the Cross-Validation method. As a result, the optimal parameters for "C" and "Solver" are 0.01 and 'newton-cg' respectively with the highest accuracy 39.63%.

Test and Discussion

The optimal model using the set of parameters C=0.01, Solver=’newton-cg’ yields an average training accuracy of 39.19% and a test set accuracy of 39.77%. Thus, training set produce slightly higher accuracy than the test set.

The optimal model is generally underfitting because, from the distribution of true target and prediction, most of the predictions made are the salary bucket of 0 (0-9,999), 10 (100,000-124,999), 12 (150,000-199,999) and 14 (>300,000). Only a few predictions were correct for other salary buckets which suggests that the model may have oversimplifying assumptions about the dataset.

Insights:

- To avoid underfit model, the number of features should be increased so the hypothesis space can be expended.
- Improving the quality of original dataset. During the data cleaning stage, most of the original feature columns were dropped due to an extremely feature columns were dropped due to an extremely low response rate. And also, some missing values were replaced by mode value, which harms the quality of the dataset and make the dataset more difficult to be differentiated by the regression algorithm.
- Increasing model complexity during the process of feature selection. Simply select the most important features may not be sufficient to develop a machine learning model with appropriate complexity. Also, apply feature engineering to the existing features helps expand the hypothesis space by adding new features based on existing features.