ASSIGNMENT 2

MIE 1624 Introduction to Data Science and Analytics

Name: Jiani Jia
Student ID: 1002226245

Data Cleaning

1. Drop columns include "OTHER_TEXT" and “TEXT”: “OTHER_TEXT” columns indicate whether participants fill the Other options or not. -1 means select the option that the survey provided, otherwise participants fill other answers by themselves. "TEXT" columns appear at "Q14_Part_TEXT", they are indicated the choice that participants choose in Q14. No useful information contains inside these columns. Remaining 219 features in the dataframe.
2. Select single choices columns and count how many Null in each of them. There are total 17 columns and none of them have more than 20% Null Answer, do not drop any of them.
3. Fill in the Null: among the 17 single choice columns, Q11, Q14 ,Q15, Q19, Q22, Q23 need to fill Null using mode() function.
4. Drop the first row: it contain questions corresponding to each column.

Data Encoding

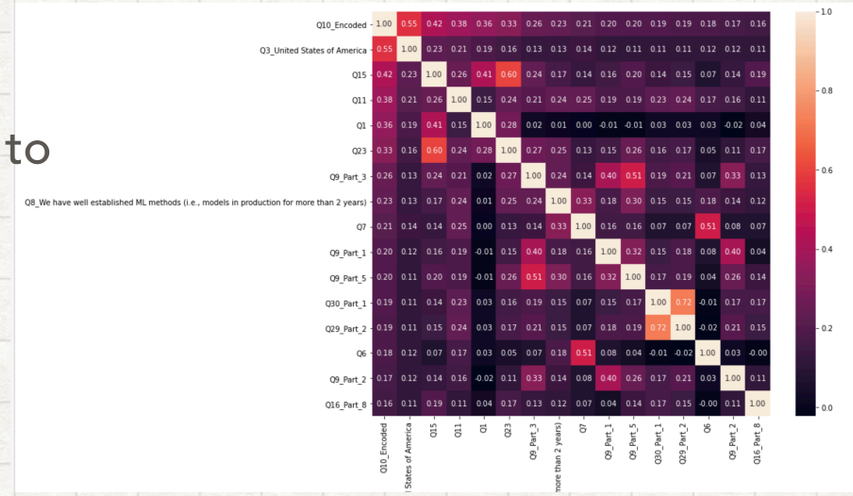
1. Label encoding: encode the ordinal categorical to an integer that can represent the order between each categorical. I apply label encoding to Q4 education level.
2. One hot encoding: for categorical variables where no ordinal relationship exists, the one hot encoding should be applied. Q2, Q3, Q5, Q8, Q14, Q19, I choose to use one hot coding. After one hot encoding, resulting 97 dummy columns and drop the original 6 columns, the remaining columns are 310.
3. Encoding data in range: for the data in range, Q1, Q6, Q7, Q11, Q15, Q22, Q23 , I choose to use the average for each range to represent each of them.
4. Encoding for multiple choice question: first select the all multiple choice questions, which are the columns name include “Part”. If there is a string exist in the columns, means this options has been chosen, so I use 1 to cover it; otherwise, give 0 to the space to indicate participants not select this option.

Exploratory Data Analysis

1. Box plot for education vs. salary: the median of average salary is increasing with the level of education. Among these seven educational options, Doctoral degree has the highest upper quartile and median, Master degree is the second-highest and Bachelor degree has the lowest median.
2. Box plot for age vs. salary: salary increases with age increasing, 60-69 age group is a peak that has the largest upper quartile and largest median. Because with the age grows, the people will have more working experience and work in a big company so they earn more.
3. Box plot for money spend on machine learning vs. salary: people spend more money on machine learning will earn more. This may because when a person earn more money, he/she will choose to spend more on machine learning, like buy an expensive computer, or dataset.

Correlation Plot

Correlation plot shows the 15 features that has the highest correlation to yearly salary. High correlation means there are strongest relationship between these columns and salary. The highest correlation is 0.55, the corresponding feature is Q3_United States of American.



Feature Selection Algorithm

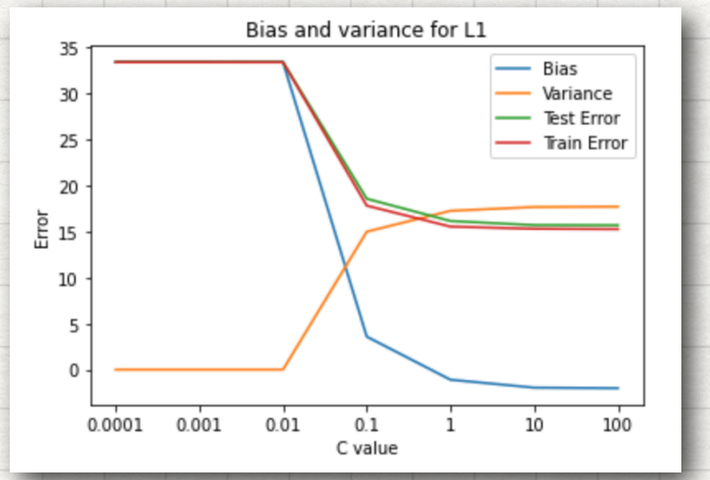
- Feature engineering is the process of extract features from raw data. It increases the predictive power of machine learning algorithms by extracting features from raw data that help facilitate the machine learning process and improve the performance of machine learning algorithms.
- Lasso for feature selection, it is one of regularized regression method, which consists in adding a penalty to the different parameters of the machine learning model to reduce the freedom of the model and in other words to avoid overfitting. Lasso or L1 has the property that is able to shrink some of the coefficients to zero. Therefore, that feature can be removed from the model.
- After Lasso, 29 features are selected. Combine with the correlation result, I choose ten features to do the machine learning. There are: Q1, Q7, Q9_Part_1, Q11, Q15, Q9_Part_3, Q3_United States of America, Q8_We have well established ML methods, Q6, Q23.

Model Implementation

Result table for L1 penalty model by verifying C value

	Average accuracy(%)	Accuracy variance(%)	Bias	Variance	Train Error	Test Error
0.0001	31.382581	9.683247	33.381690	0.000000	33.381903	33.381690
0.0010	31.382581	9.683247	33.381690	0.000000	33.381903	33.381690
0.0100	31.382581	9.683247	33.381690	0.000000	33.377099	33.381690
0.1000	33.706452	6.289342	3.575870	14.974788	17.802939	18.550658
1.0000	33.226452	4.881136	-1.102298	17.239427	15.527872	16.137129
10.0000	33.307097	6.449550	-1.963184	17.653171	15.290511	15.689987
100.0000	33.307097	6.449550	-2.024150	17.696537	15.248251	15.672387

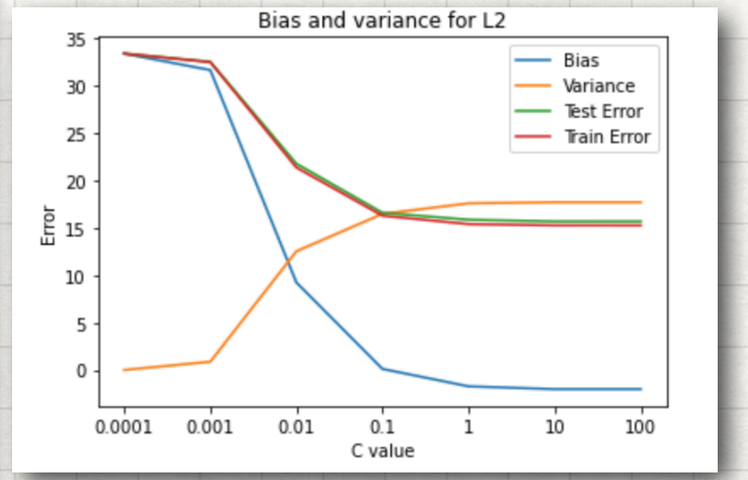
	Probability
0-9,999	0.304113
10,000-19,999	0.104458
20,000-29,999	0.091846
30,000-39,999	0.064074
40,000-49,999	0.053947
50,000-59,999	0.040491
60,000-69,999	0.038310
70,000-79,999	0.059859
80,000-89,999	0.029452
90,000-99,999	0.029595
100,000-124,999	0.069996
125,000-149,999	0.042334
150,000-199,999	0.032142
200,000-249,999	0.021523
>250,000	0.017860



Result table for L2 penalty model by verifying C value

	Average accuracy(%)	Accuracy variance(%)	Bias	Variance	Train Error	Test Error
0.0001	31.382581	9.683247	33.381690	0.000000	33.381903	33.381690
0.0010	31.703226	7.960674	31.647224	0.868402	32.481618	32.515626
0.0100	33.385161	6.047401	9.236274	12.525081	21.363305	21.761355
0.1000	33.386452	7.407071	0.109894	16.473506	16.277453	16.583400
1.0000	33.307097	5.553550	-1.729437	17.585024	15.390429	15.855587
10.0000	33.307097	6.449550	-2.024150	17.696537	15.257859	15.672387
100.0000	33.307097	6.449550	-2.024150	17.696537	15.254034	15.672387

	Probability
0-9,999	0.316285
10,000-19,999	0.107510
20,000-29,999	0.083924
30,000-39,999	0.057168
40,000-49,999	0.058307
50,000-59,999	0.058089
60,000-69,999	0.048683
70,000-79,999	0.041636
80,000-89,999	0.034727
90,000-99,999	0.030151
100,000-124,999	0.061147
125,000-149,999	0.039654
150,000-199,999	0.032801
200,000-249,999	0.011380
>250,000	0.018536



- From the output table of the algorithm, it shows approximately 30% probability that the prediction result will fall into the 0-9999 USD salary bucket.
- Tradeoff in complexity means there is a tradeoff between bias and variance, the algorithm can not be more complex and less complex at the same time. The optimal model complexity is occur at the intersection of bias and variance. From the bias and variance trade off graphs for L1 and L2, the intersection for bias and variance for L1 and L2 are approximately 0.09 and 0.009, at these times, L1 will result a smaller test error, so I choose the model with hyperparameter C=0.09 penalty=l1 solver=saga, the accuracy is 0.337.

Model Tuning

- After model implementation, I did model tuning applying the grid search.
- Grid search score: 0.36156
- Grid search hyper parameter: C=0.351, solver=saga, penalty=l1, max_iter=100

Performance Measures

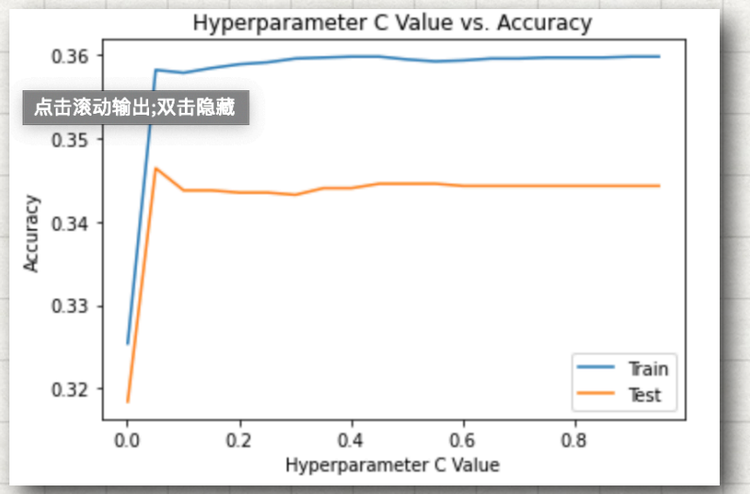
clasification report:

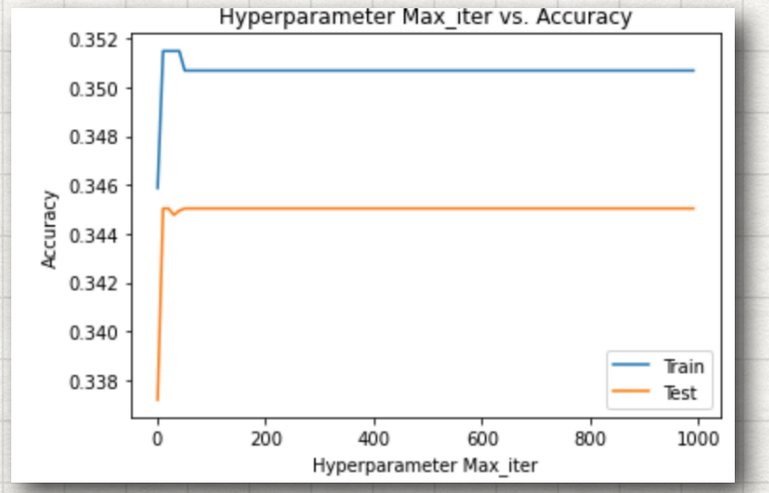
	precision	recall	f1-score	support
0-9,999	0.46	0.91	0.61	129
10,000-19,999	0.06	0.02	0.03	44
20,000-29,999	0.00	0.00	0.00	29
30,000-39,999	0.00	0.00	0.00	24
40,000-49,999	0.00	0.00	0.00	18
50,000-59,999	0.22	0.11	0.15	18
60,000-69,999	0.00	0.00	0.00	20
70,000-79,999	0.25	0.05	0.09	19
80,000-89,999	0.00	0.00	0.00	9
90,000-99,999	0.00	0.00	0.00	8
100,000-124,999	0.23	0.48	0.31	23
125,000-149,999	0.09	0.25	0.13	8
150,000-199,999	1.00	0.07	0.12	15
200,000-249,999	0.33	0.17	0.22	6
>250,000	0.00	0.00	0.00	5
accuracy			0.37	375
macro avg	0.18	0.14	0.11	375
weighted avg	0.25	0.37	0.26	375

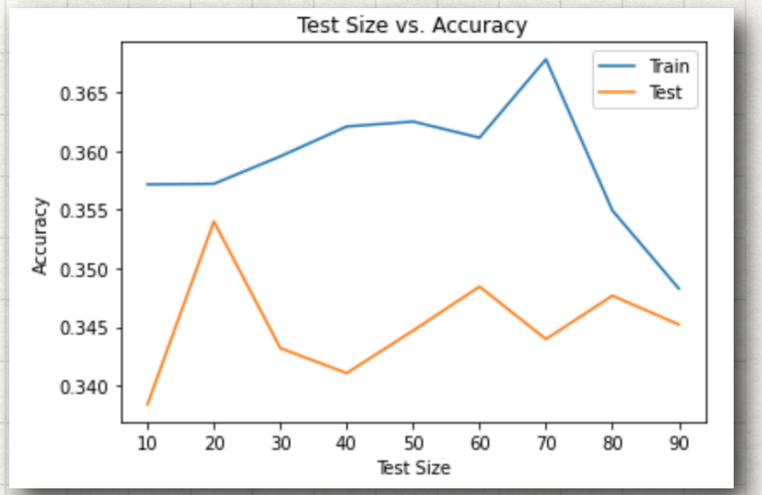
- Using the model hyper parameters get from grid search, the accuracy is 0.37, which is higher than the model in model implement.
- The optimal model is LogisticRegression(C=0.351, solver=saga, penalty=l1, max_iter=100)

Test & Discussion

- For optimal model, training accuracy is 0.3507 and test accuracy is 0.345







Test & Discussion

- The first plot is keeping the max_iter as 100, tuning C value and the second plot is keeping C as 0.35 changing max_iter. Both of the plot shows the train and test accuracies are not improved, both of them are below 40%.
- Since grid search is already try all metrics combinations, so changing hyper parameter will not improve the training and test accuracy.
- For changing training or test set size will give a little bit improve, but the overall accuracy for both training and test accuracies are below 40%, witch are quite low.
- Both training and test accuracies are low means the model is underfitting.

Reason for Underfitting:

- One reason is because the model or the algorithm does not fit the data well enough, it usually happens when we have less data to build an accurate model. If more features added into training and test process, the accuracy will improved.
- Another reason is the logistic regression can not fit to our multi-class classification dataset. Since logistic regression is a form of binary regression, it not work will to classify more than two classes.
- It also may because the unbalance dataset, most participants earn the salary is between 0 to 20,000 USD, so that the model will predict more in that range, it will also cause the low accuracy.