MIE1624 Assignment1

WeiinLiu 1005482826

Outline

- Data Description
- Explorative analysis
- Feature Engineering
- Model Implementation
- Model Results

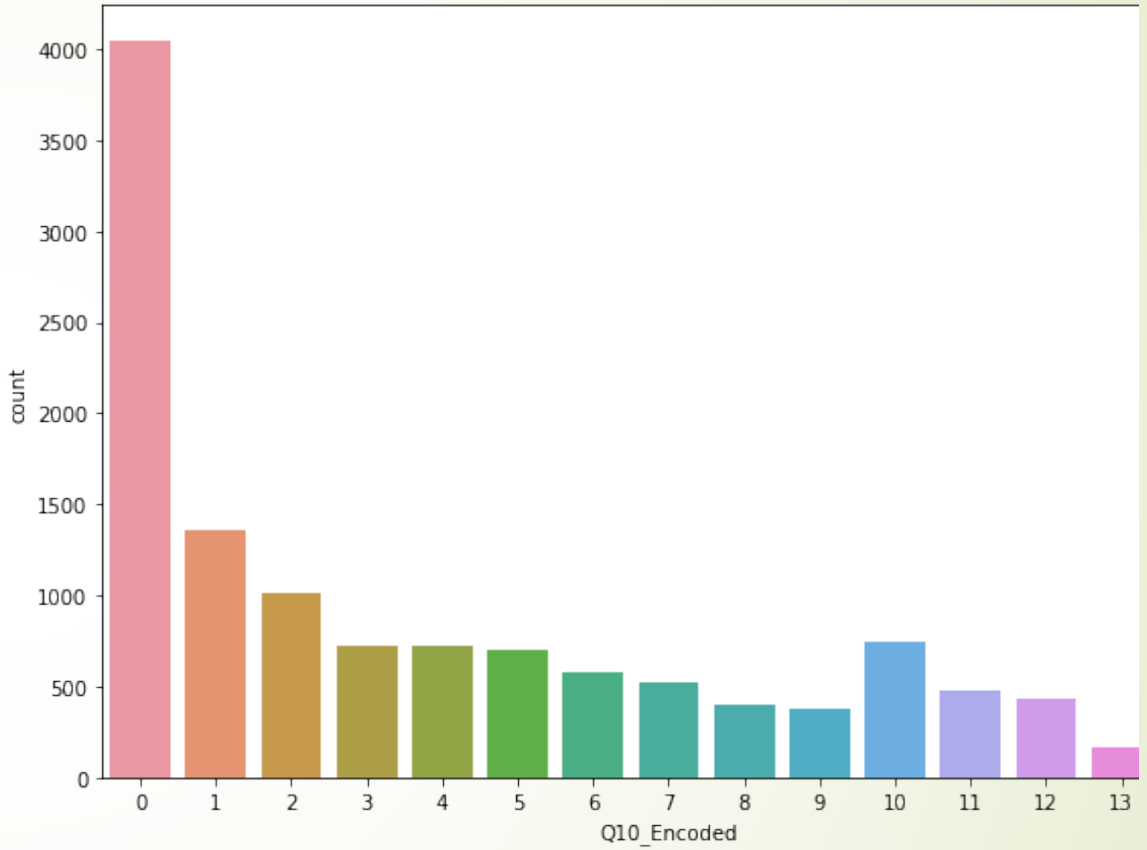
Data Description

- Range Index: 12497 entries, 0 to 12496
- Columns: 248 entries, Time from Start to Finish (seconds) to Q10_buckets
- memory usage: 23.6+ MB

Explorative Analysis

- The distribution of labels

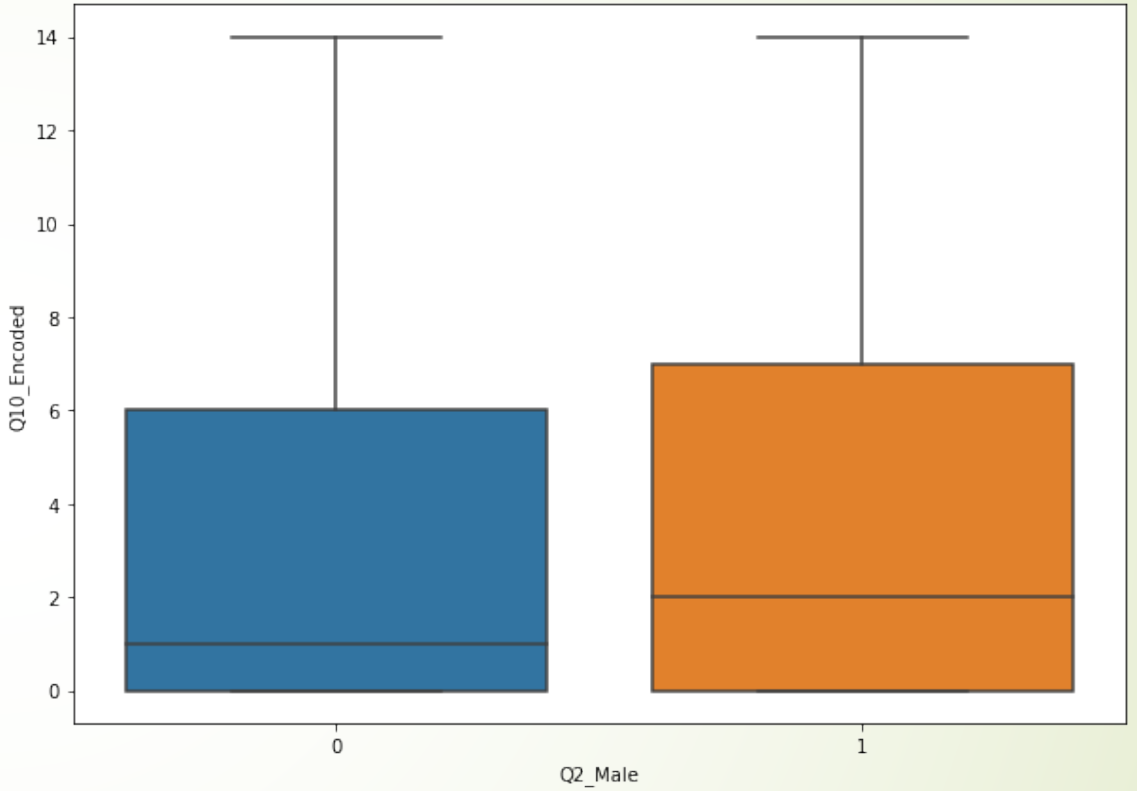
In the first band of salary, it has a great proportion. And as the salary increases, the count decreases.



Explorative Analysis

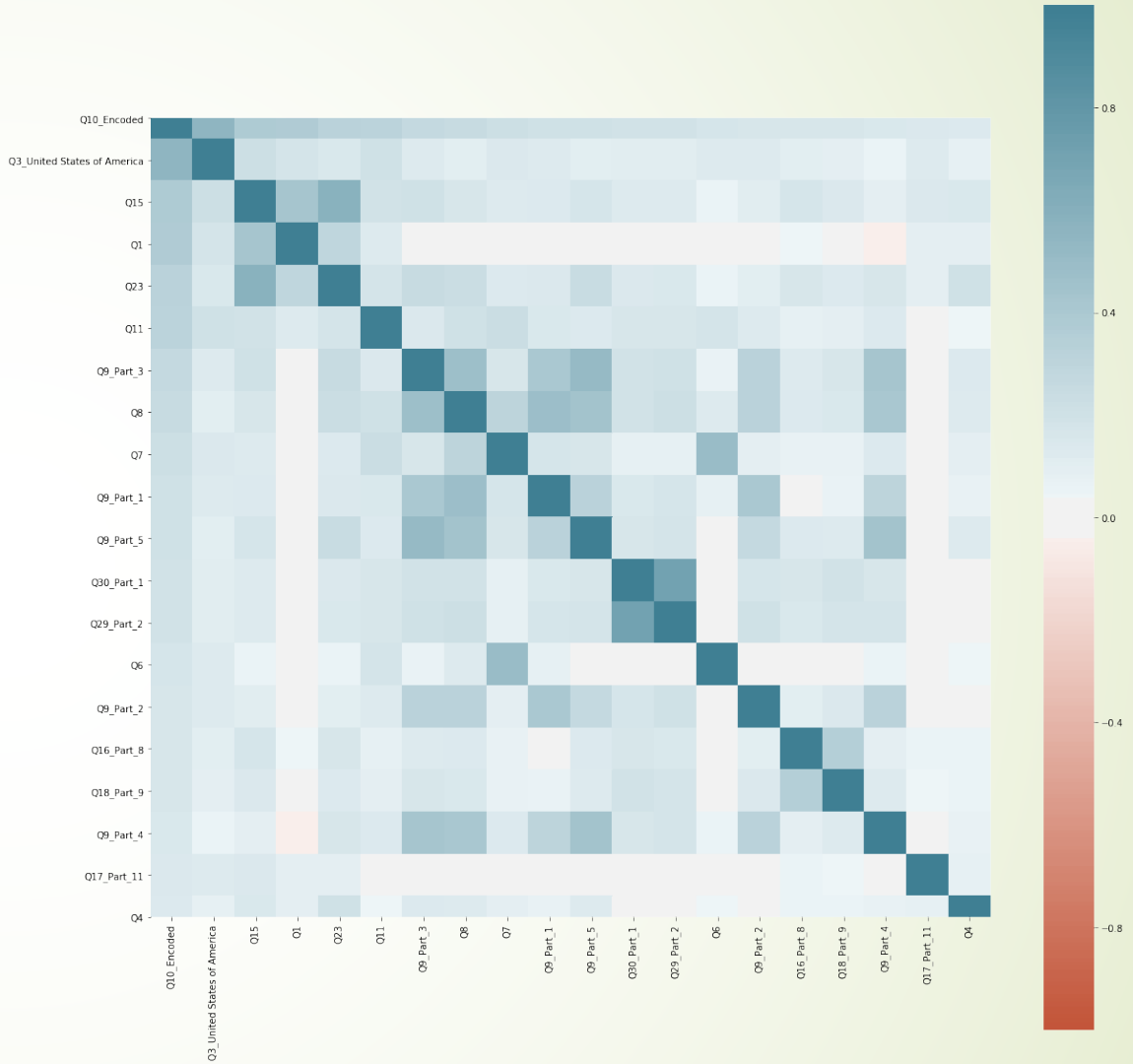
- Gender vs Salary

Male works tend to have higher salary in averge.



Feature Engineering

- Fill missing value with medium value of the column. The columns includes Q11, Q14, Q15, Q19, Q22, Q23.
- Drop useless columns includes Q10, Q10_buckets, Time from Start to Finish (seconds) and columns with “TEXT”.
- Convert categorical values with one-hot encoding.
- Delete features with small variance
- Select features with large mutual info



Model implementation

- Logistic regression model is chosen.
- 10-fold cross validation is used to evaluate the model
- GridSearchCV is applied for selecting the optimal parameters
- The optimal parameters are

  {'C': 0.05, 'solver': 'newton-cg'}

Model Results

- The accuracy could reach 35%
- The performance in predicting 4,9 is pool
- The performance in predicting 0 is acceptable.

Detailed classification report on test data:

	precision	recall	f1-score	support
0	0.43	0.95	0.59	1194
1	0.11	0.03	0.05	405
2	0.09	0.04	0.06	281
3	0.10	0.03	0.04	224
4	0.09	0.04	0.05	221
5	0.04	0.01	0.02	218
6	0.07	0.02	0.03	173
7	0.22	0.04	0.07	159
8	0.11	0.01	0.02	113
9	0.00	0.00	0.00	125
10	0.15	0.26	0.19	221
11	0.21	0.14	0.17	150
12	0.28	0.22	0.25	143
13	0.00	0.00	0.00	56
14	0.17	0.06	0.09	67
accuracy			0.35	3750
macro avg	0.14	0.12	0.11	3750
weighted avg	0.22	0.35	0.24	3750