MIE 1624
ASSIGNMENT 1
SALARY CLASSIFICATION

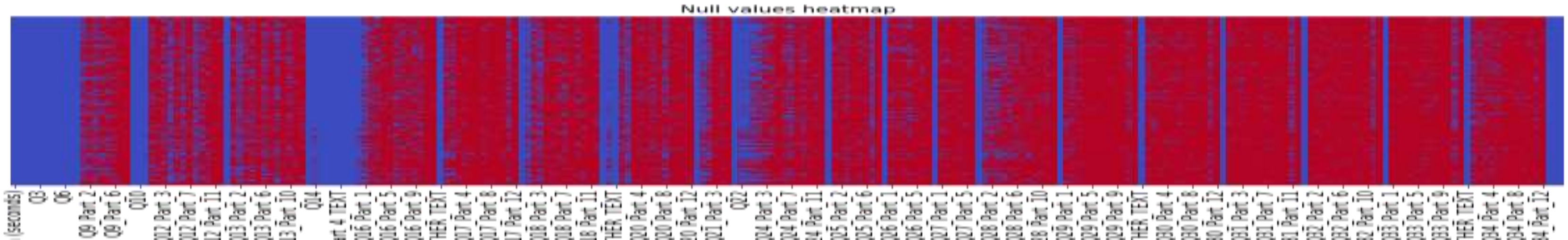
Anshul Verma
1004730703

Date: 28th. February. 2020

Q1. Data Cleaning

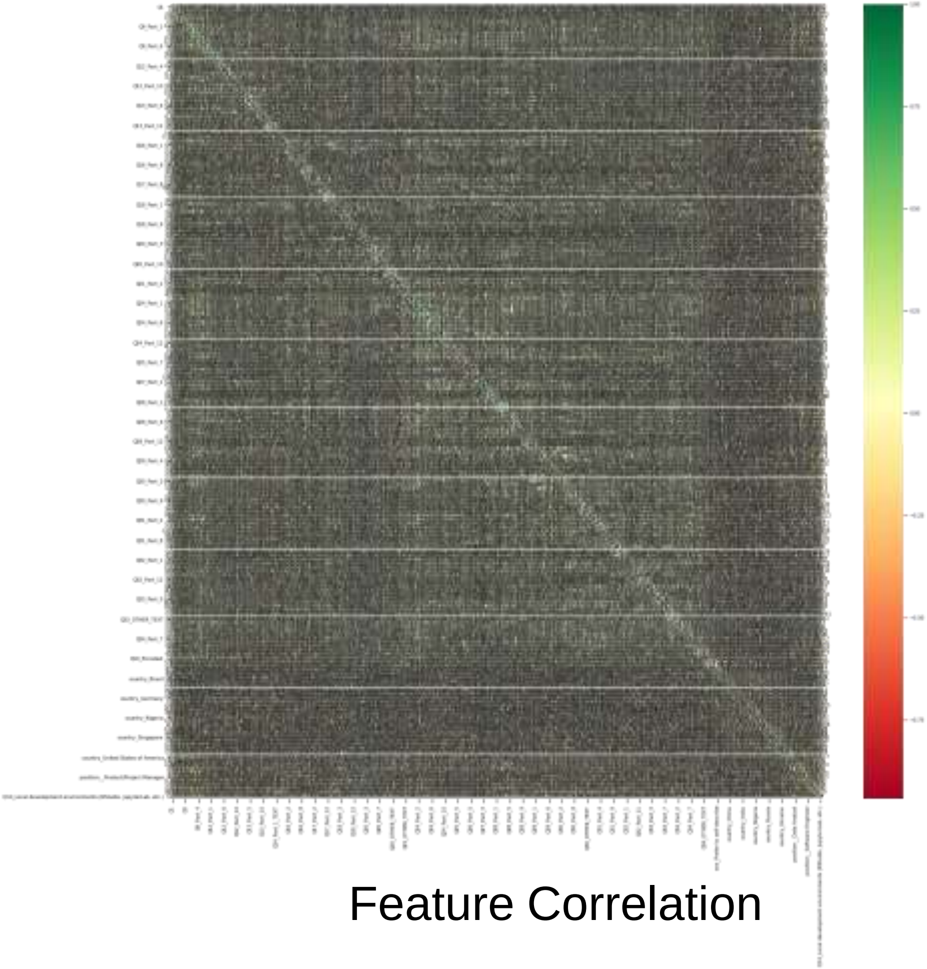
- Most of the features in the data-set where binary either they had an entry or were null. So I encoded all these columns in 0-1 form.
- Then columns like sex, country, job-positions were treated as nominal categories and one-hot encoding was used for these types of columns.
- Other columns like education level, age, experience, Use of ML etc. were treated as ordinal categorical data and were encoded in a way that preserves the order.
- After encoding all the major columns there were some Nan values still present in the data-frame so the rows with Nan were removed from the dataset (I did this because the number of Nans were very small)

Heat map of data-frame without cleaning



Q2 & 3. Exploratory Analysis & Feature Selection

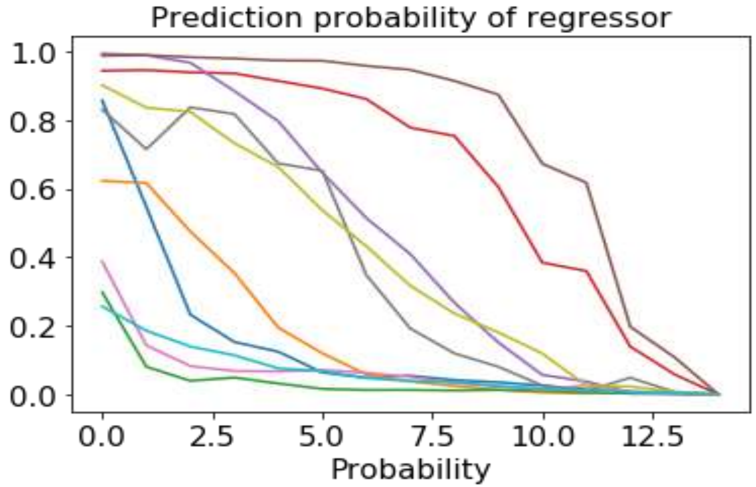
- For Exploratory data analysis the final salary bucket vs features which seem more influential is plotted and presented in the jupyter notebook.
- Correlation between features was used for feature selection
- Based on Correlation top 200 features were selected and used for the purpose of logistic regression.
- Finally the features and targets were seperated and features were further used to build an ordinal logistic regression model.

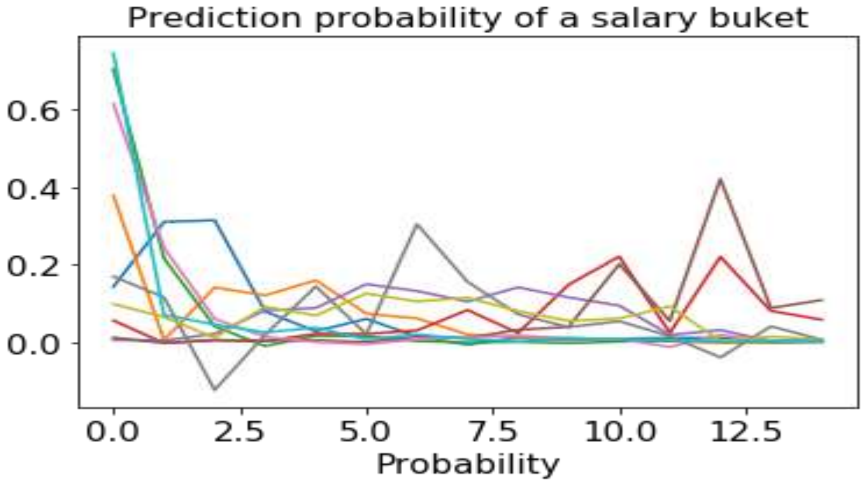


Feature Correlation

Q4. Model Implementation

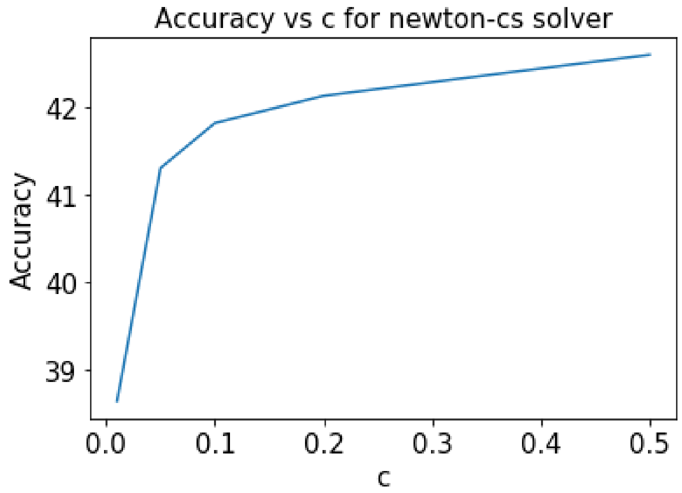
- 14 different binary linear regressor were used to classify all the classes.
- The classification for every class was of the form 0(<=salary_bucket) and 1(>salary_bucket).
- Using probabilities for this mixed classification probabilities of each class was identified.
- Then Based on the each classes probability predictions were made.
- The models were then used for 10-fold cross validation to make sure that the model doesn’t overfit. (Making sure standard deviations of accuracy is low)

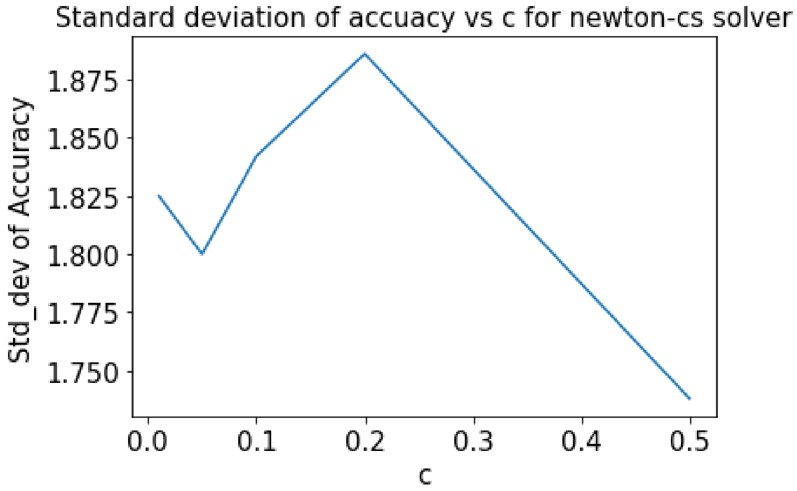




Q5. Model Tuning

- Grid search on some possible hyper-parameter list was performed.
- There can be more hyper-parameter like regularize type, number of iterations. But for the purpose of assignment only L2 regularizer was used and max_iter was set to 100.
- The final selected model had a 10-fold average accuracy of 42.6% with standard deviation 1.73%.
- This model was finally used to make predictions on the test set.





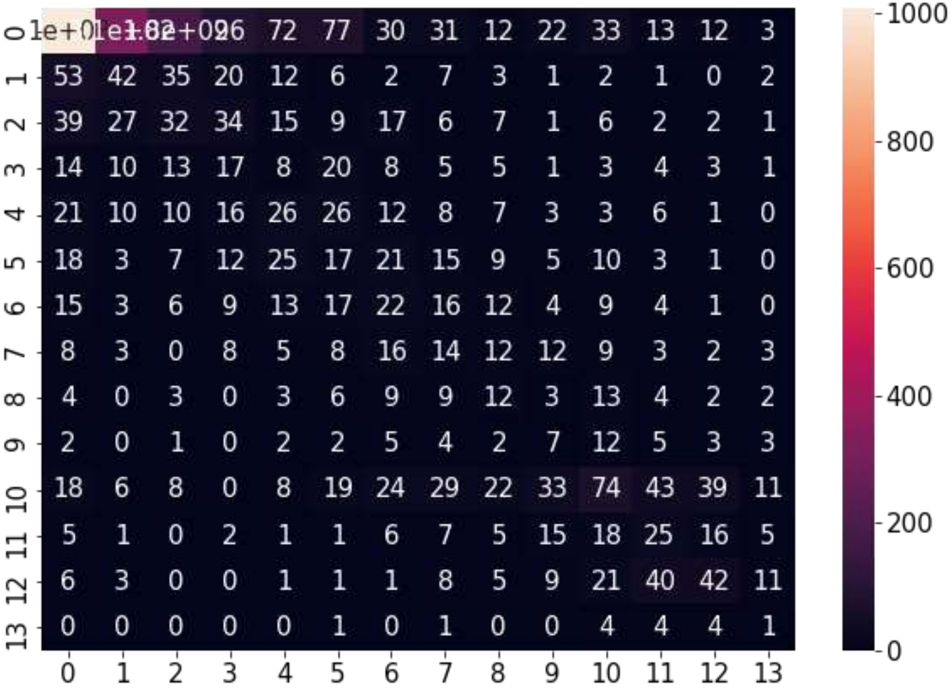
Q4. Testing

- The final selected model was then used to test the points in the test-set.

- The model was found to be 35.9% accurate which is pretty close to training error(42.6%). This shows that generalization was decent as test data was not seen while training

- The accuracy is not very high because of a lot of classes and simplicity of the ordinal logistic regression model.

- Using some other classification approach can increase the



	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	1e+0	1e+02	1.8e+02	96	72	77	30	31	12	22	33	13	12	3
1	53	42	35	20	12	6	2	7	3	1	2	1	0	2
2	39	27	32	34	15	9	17	6	7	1	6	2	2	1
3	14	10	13	17	8	20	8	5	5	1	3	4	3	1
4	21	10	10	16	26	26	12	8	7	3	3	6	1	0
5	18	3	7	12	25	17	21	15	9	5	10	3	1	0
6	15	3	6	9	13	17	22	16	12	4	9	4	1	0
7	8	3	0	8	5	8	16	14	12	12	9	3	2	3
8	4	0	3	0	3	6	9	9	12	3	13	4	2	2
9	2	0	1	0	2	2	5	4	2	7	12	5	3	3
10	18	6	8	0	8	19	24	29	22	33	74	43	39	11
11	5	1	0	2	1	1	6	7	5	15	18	25	16	5
12	6	3	0	0	1	1	1	8	5	9	21	40	42	11
13	0	0	0	0	0	1	0	1	0	0	4	4	4	1

Confusion Matrix of Test Prediction