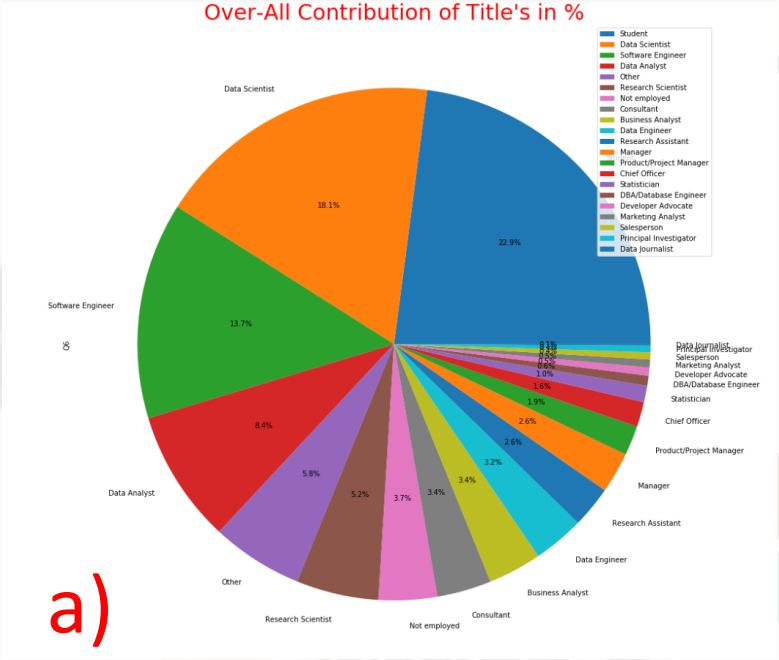
# Project 1
# Data Science Salary Classification Problem

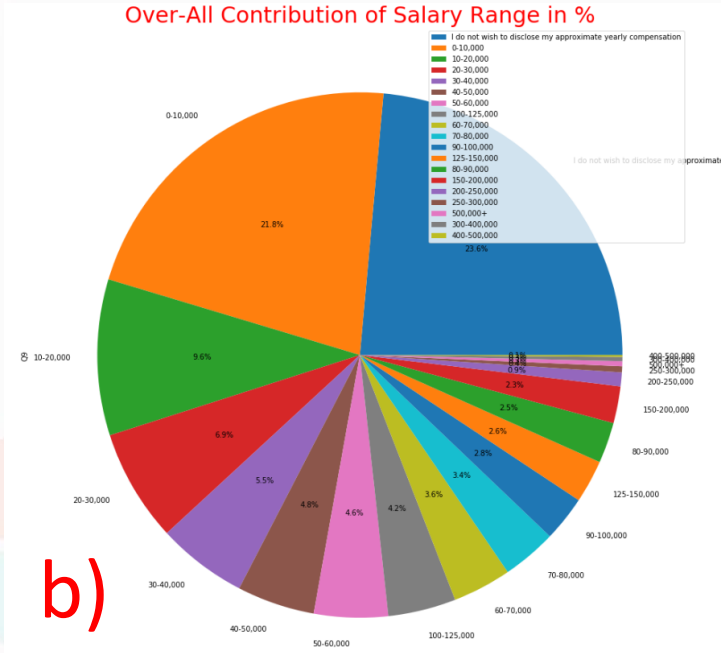

Saket Thavanani

Student Number-1005643145

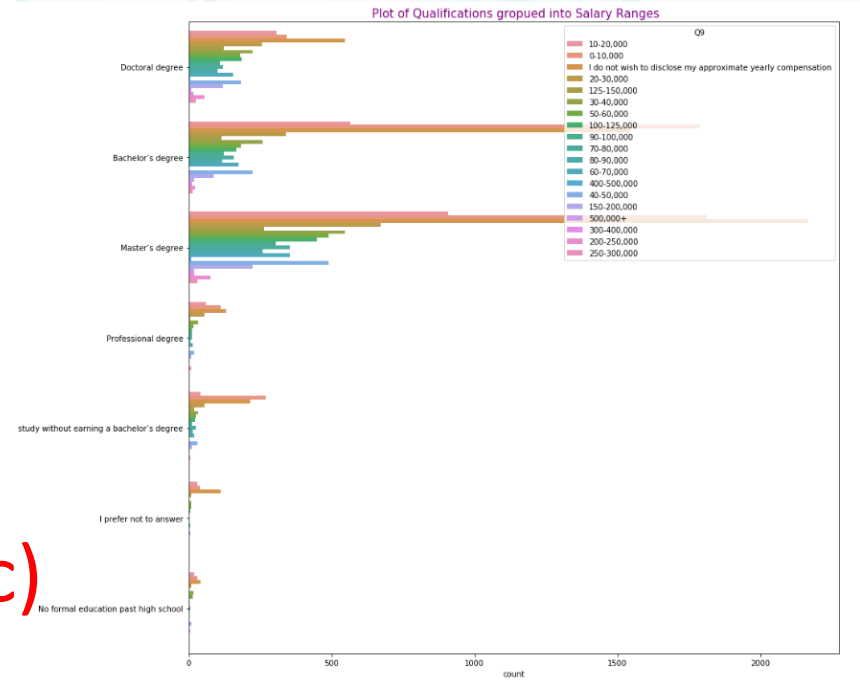
# Exploratory data analysis




a)




b)




c)

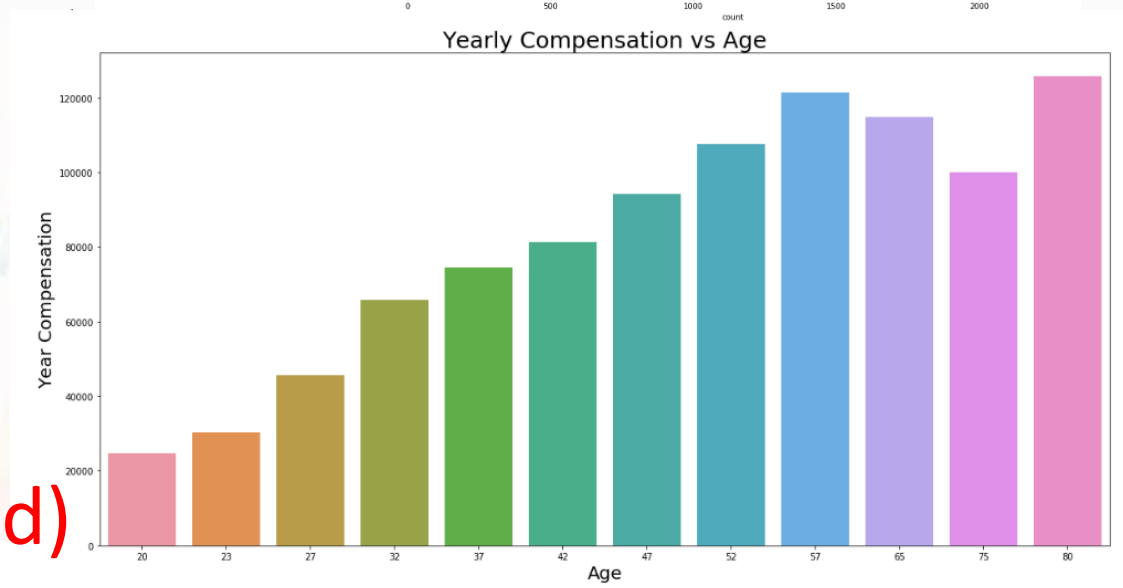
**Inference**-

a) Data Science is one of the most attracted field today, peoples from different professions are moving towards Data Science i.e. various persons from other fields like Software Engineer, Student, Research Assistant, Chief Officer, Manager, Research Scientists, etc. are moving towards data science as a career option.

b) We can see from the above plot that most of the people have their yearly compensation lying in the range of 0-10,000k. Approximately,21.8% people have their compensation in range of 0-10,000k

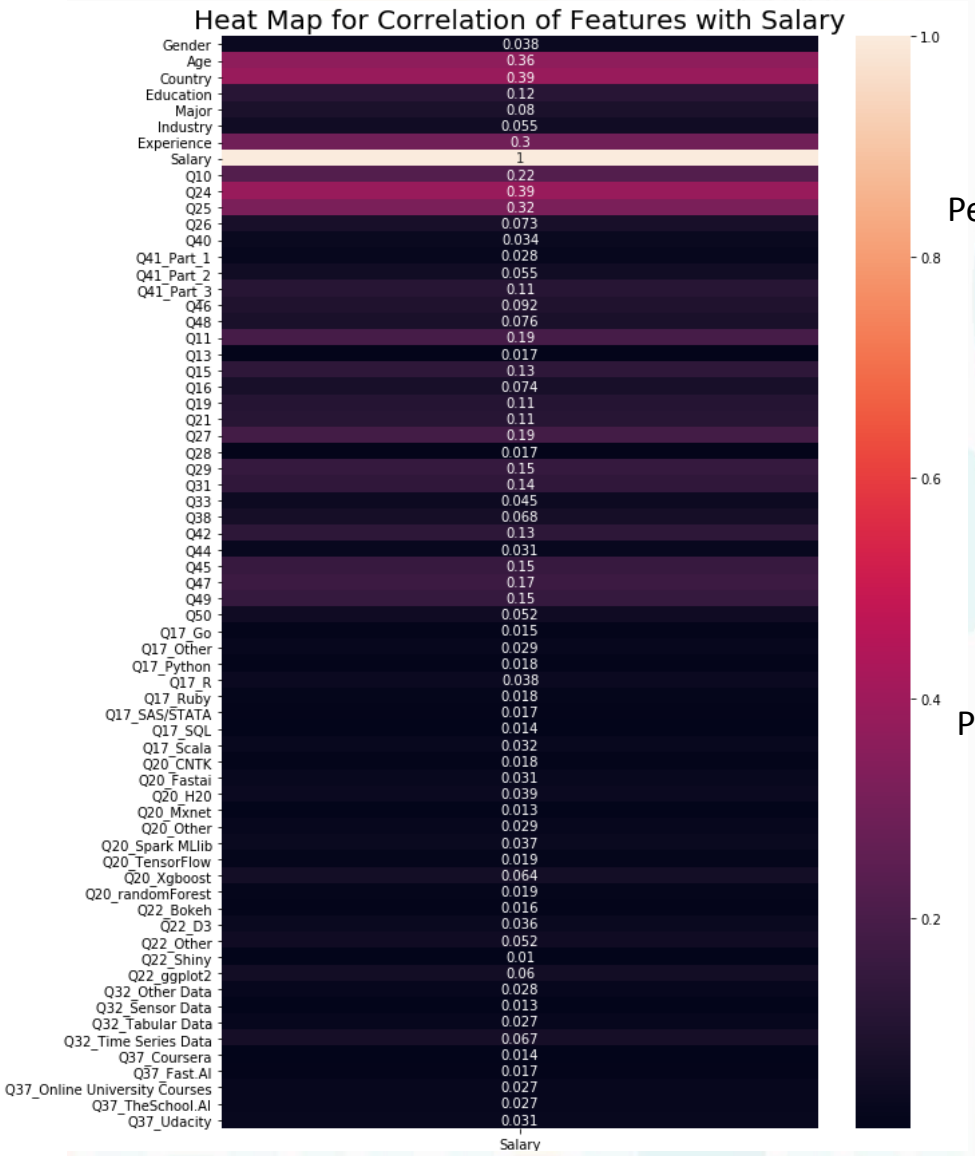
c-) We can see from the above plot that. People who have Doctoral's and Masters degree have higher yearly compensations as well as compared to bachelors and normal college degree

d-) We can infer from the above plot that age is in high correlation with the yearly compensation ,i.e People who are elder they usually have high yearly compensation. Hence, Age can be a very important feature in predicting Salary.

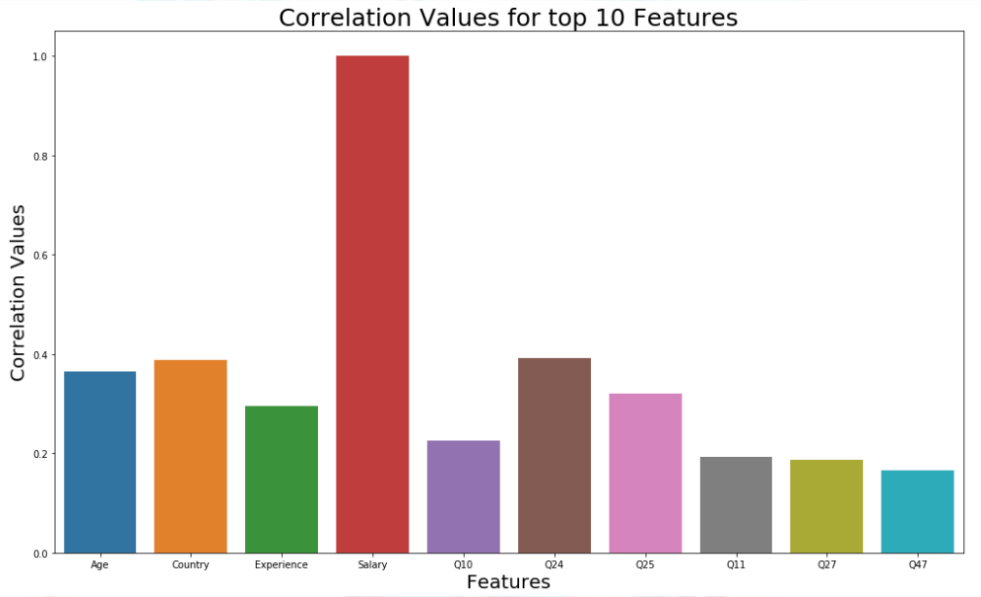



d)

# Exploratory data analysis




Top 10 features
Pearson correlation coefficients > 0.16




Filtered 66 features which have a good correlation with the output target Salary Ranges

Feature Selection
Pearson correlation coefficients > 0.01

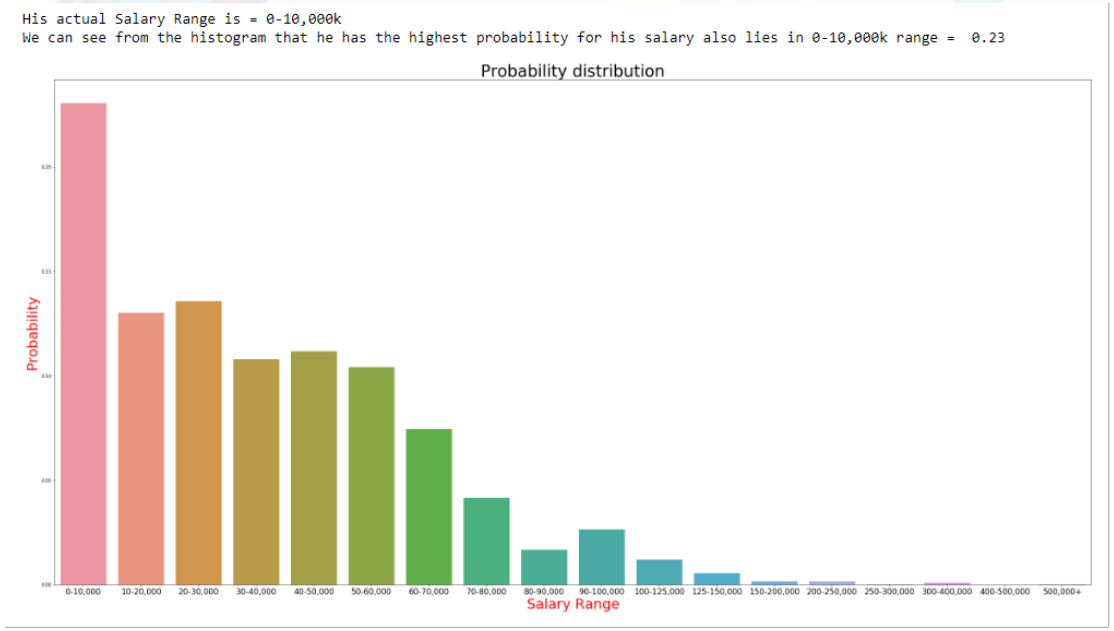
```
Index(['Gender', 'Age', 'Country', 'Education', 'Major', 'Industry',
       'Experience', 'Salary', 'Q10', 'Q24', 'Q25', 'Q26', 'Q40', 'Q41_Part_1',
       'Q41_Part_2', 'Q41_Part_3', 'Q46', 'Q48', 'Q11', 'Q13', 'Q15', 'Q16',
       'Q19', 'Q21', 'Q27', 'Q28', 'Q29', 'Q31', 'Q33', 'Q38', 'Q42', 'Q44',
       'Q45', 'Q47', 'Q49', 'Q50', 'Q17_Go', 'Q17_Other', 'Q17_Python',
       'Q17_R', 'Q17_Ruby', 'Q17_SAS/STATA', 'Q17_SQL', 'Q17_Scala',
       'Q20_CNTK', 'Q20_Fastai', 'Q20_H20', 'Q20_Mxnet', 'Q20_Other',
       'Q20_Spark MLlib', 'Q20_TensorFlow', 'Q20_Xgboost', 'Q20_randomForest',
       'Q22_Bokeh', 'Q22_D3', 'Q22_Other', 'Q22_Shiny', 'Q22_ggplot2',
       'Q32_Other Data', 'Q32_Sensor Data', 'Q32_Tabular Data',
       'Q32_Time Series Data', 'Q37_Coursera', 'Q37_Fast.AI',
       'Q37_Online University Courses', 'Q37_TheSchool.AI', 'Q37_Udacity'],
      dtype='object')
```

# Model Implementation

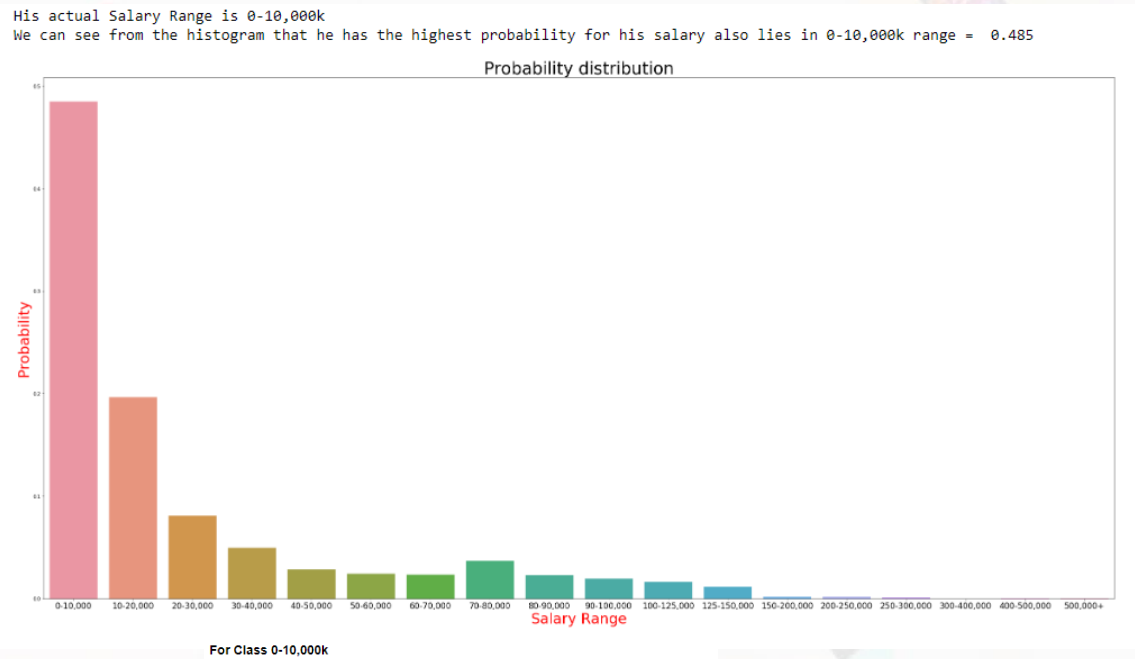
## Performance on Training and Test set

- First, the cross validation has been done on the training set.
- I will take two random persons 1 from training and other from testing set and see how well is my model predicting their probability distribution for salary ranges.
- For my case, lets chose that random person to be the 8th person from training as well as testing set.

### Training Set




### Test Set




### Inference-

Recall, Precison , F1score have come out be very good in prediction of 0-10,000k Class
My model is working pretty well in predicting 0-10,000k class
Basesd upon Confusion matrix –
Recall_10k=79.69%
Precision_10k=34%
F1-Score_10k=47.6%

**For Class 0-10,000k**

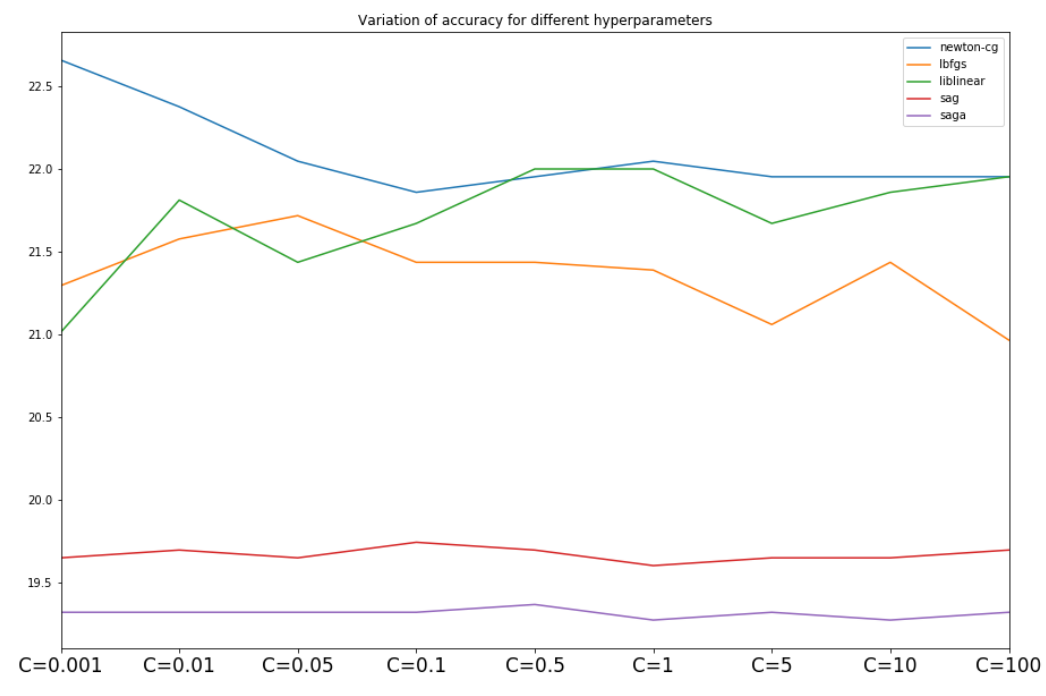
```
TP=conf_df.iloc[0,0]
FN=conf_df.iloc[0,1:].sum()
FP=conf_df.iloc[1:,0].sum()

Recall_10k=TP/(TP+FN)
Precision_10k=TP/(TP+FP)

print("Recall for 0-10,00k =",Recall_10k)
print("Precision for 0-10,00k =",Precision_10k)

F1_score=(2*(Recall_10k)*(Precision_10k))/(Precision_10k + Recall_10k)

print("F1 Score for class 0-10,000k = ",F1_score)
```

```
Recall for 0-10,00k = 0.796923076923077
Precision for 0-10,00k = 0.3394495412844037
F1 Score for class 0-10,000k =  0.4761029411764706
```

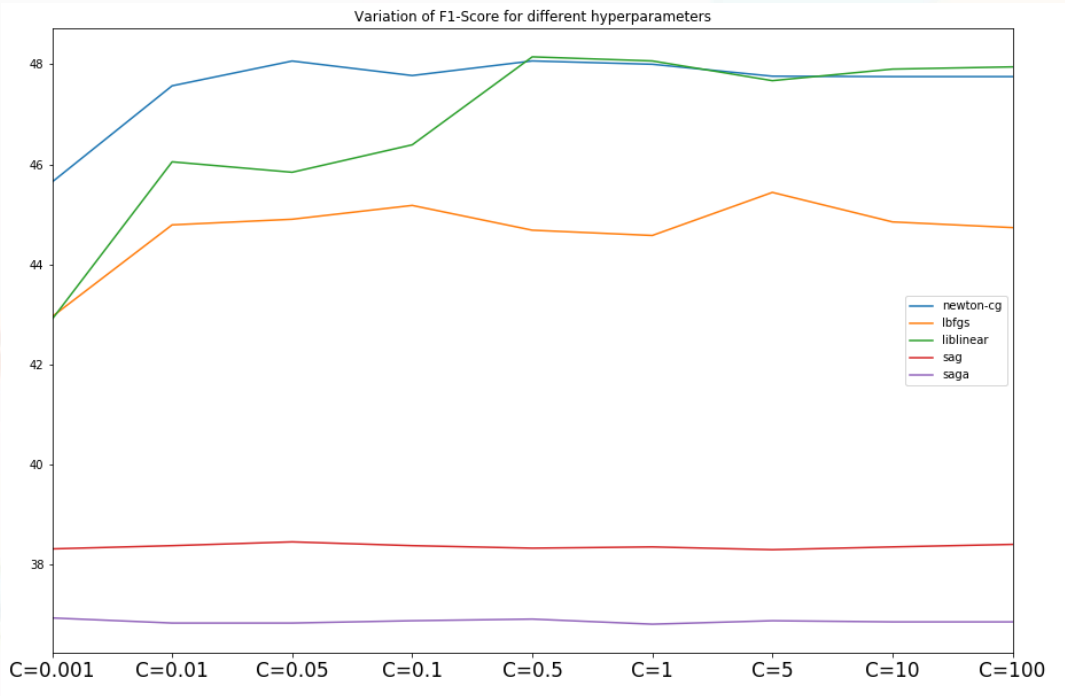
# Hyperparameter Tuning: Grid Search

- For, finding the optimum value of hyperparameters I varied two hyperparameters C and solver.
- For different combinations of C value and solver, I computed Precision, Recall, F1-Score and Accuracy for all the possible combinations and plotted their variation in a graph. Then I selected the combination of hyperparameters. which gave best of these scores.

## Accuracy Variation For Different C and Solver Combination




## F1-score 10k Variation For Different C and Solver Combination




Inference-

- We can see from the above plot that best accuracy is yield for C=0.001 solver as newton-cg.
- We can see from the above plot that best F1-Score is yield for C=0.05 solver as newton-cg
- But since Accuracy gives the overall accuracy for the model comparing all the classes so i think Accuracy would be a better criterion to judge the model. Whereas, On the other hand, F1-Score, Recall and Precision for this class 0-10,00k is good but it may not be good for all the classes at the same time.
- Actually Precision, recall are bad for other classes if we compute with the confusion matrix. Hence, Accuracy would be the best criterion to judge our Model.
- But, for value of C=0.001 the training and test scores seem to get very close to each other with increasing complexity as found towards the end of results. Which leads to high bias. Hence, i opted for second best value of C=0.05.
- SO, C=0.05 will be chosen as best Hyperparameter with Solver = newton-cg.

# Testing

## Results-

- The best model with best hyperparameters achieved accuracy -22% on testing set and 25% on training set.
- If training curve has a much better score but testing curve has a lower score, i.e., there are large gaps between two curves. Then the model suffer from an over fitting problem (High Variance).
- So, if two curves are "very close to each other" and both of them but have a low score. The model suffer from an under fitting problem (High Bias)
- For our case the gap is optimum and after certain threshold for the complexity of the model the lines are becoming parallel to each other.
- This signifies that our model is neither underfit nor overfit. It is fit.
- Accuracy can be increased by these ways-

➢ Optimize other scores for all classes-F1-Score,Precision,recalll
➢ Hyperparameter Tuning–Grid Search for more parameters-C, Solver ,Penalty, max_iterations, Multiclass option.
➢ Looking for class imbalance in my data.
➢ Detailed Error Analysis

## Learning Curve