Campus Recruitment Predictions

- Pravesh Srivastava

The Recruitment Problem

- Campus Placement is becoming highly competitive everyday.
- With increase in population, it is necessary to think out of the box.
- Must work hard from start so as to earn a decent future salary and job.
- CO-Curricular do matter a lot.

The dataset

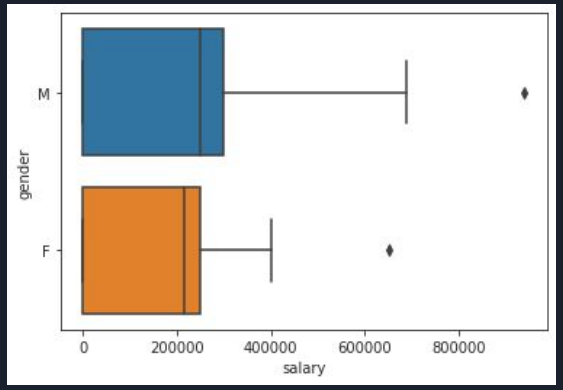
- Dataset has been downloaded from kaggle.com which is open competitive platform for data scientist. Dataset can be accessed by clicking this link.
- It consists of 215 instances nad 15 columns.
- The dataset is provided by MBA student doing MBA in business analytics at Jain University, Bangalore.

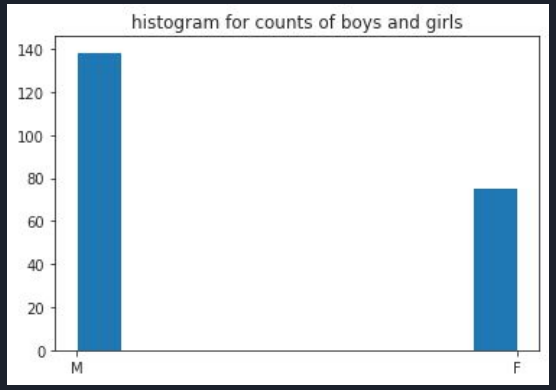
	sl_no	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	status	salary
0	1	M	67.00	Others	91.00	Others	Commerce	58.00	Sci&Tech	No	55.0	Mkt&HR	58.80	Placed	270000.0
1	2	M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28	Placed	200000.0
2	3	M	65.00	Central	68.00	Central	Arts	64.00	Comm&Mgmt	No	75.0	Mkt&Fin	57.80	Placed	250000.0
3	4	M	56.00	Central	52.00	Central	Science	52.00	Sci&Tech	No	66.0	Mkt&HR	59.43	Not Placed	NaN
4	5	M	85.80	Central	73.60	Central	Commerce	73.30	Comm&Mgmt	No	96.8	Mkt&Fin	55.50	Placed	425000.0

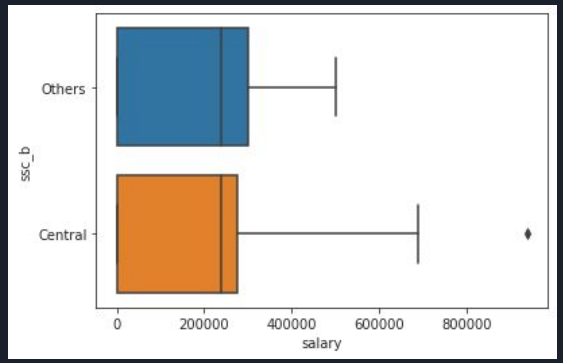
Use cases

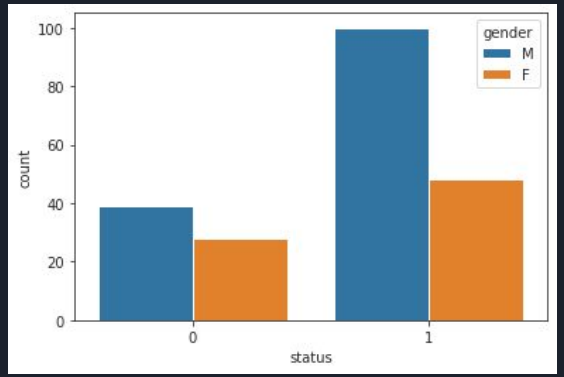
- Classification Problem for predicting placed or unplaced.
- Regression problem for the amount of salary to be received on getting placed.

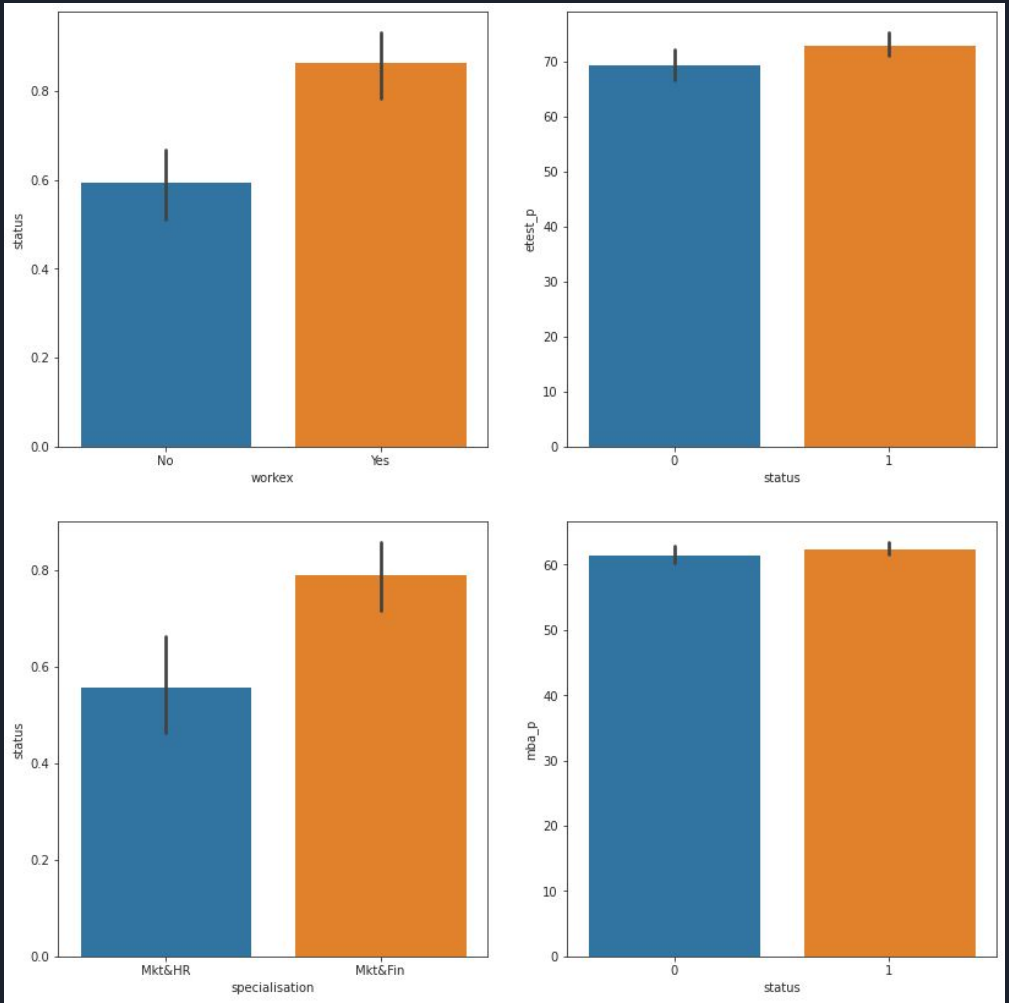
Data Exploration







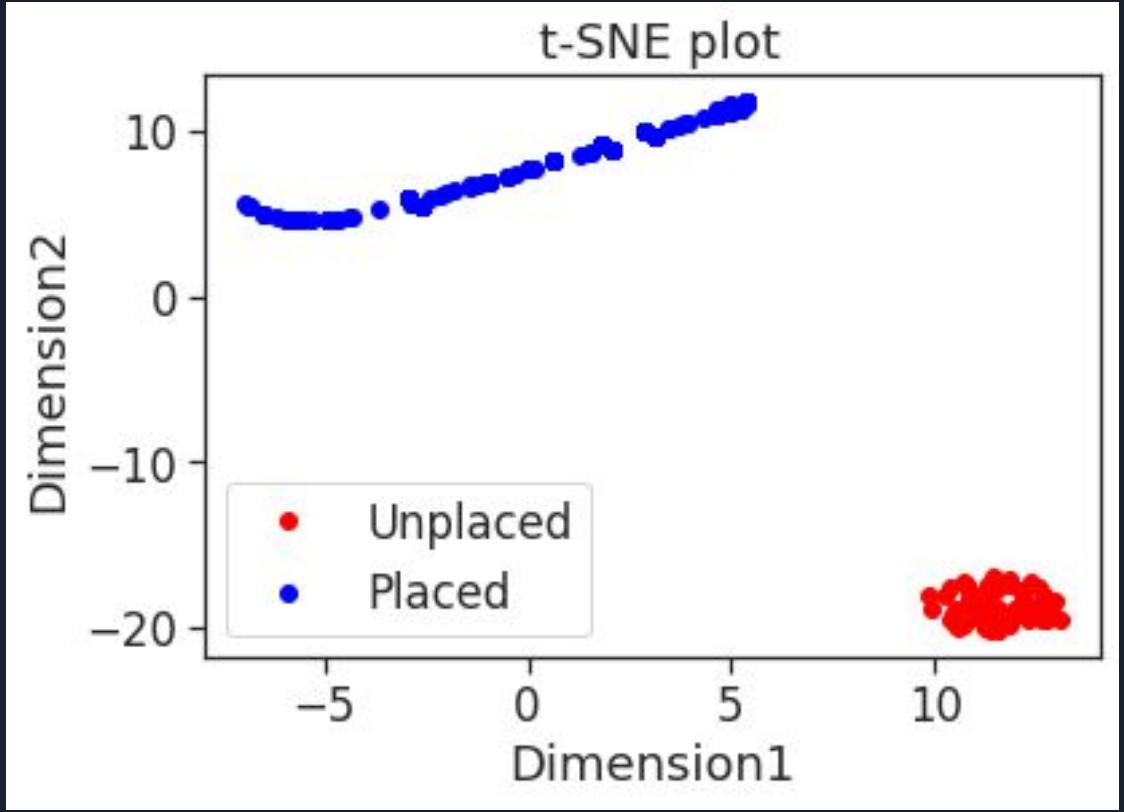



status
0.0
0.2
0.4
0.6
0.8
No
Yes
workex
etest_p
0
10
20
30
40
50
60
70
0
1
status
status
0.0
0.2
0.4
0.6
0.8
Mkt&HR
Mkt&Fin
specialisation
mba_p
0
10
20
30
40
50
60
0
1
status

Preprocessing and Feature Engineering

- Backward difference encoding of the categorical features and features containing texts.
- Dropping the features not relevant or which puts less impact on the model.
- Extracting important features.
- Using PCA/t-SNE for distribution visualization.
- Standardizing the dataset before training.

t-SNE Plot


t-SNE plot
10
0
−10
−20
Dimension2
−5
0
5
10
Dimension1
Unplaced
Placed

Classifier used and performance measure

- Classifier
  - DecisionTreeClassifier
  - RandomForestClassifier
  - XGBClassifier
  - GaussianNB
  - KNeighborsClassifier
  - MLPClassifier
- Regression
  - RandomForestRegressor
  - from xgboost import XGBRegressor
- Performance Measure
  - F1 Score
  - R^2 Score

Performance

- Classification problem
  - All the classifier got F1 score of 1.
- Regression Problem
  - R^2 score (best achieved) is 0.52 and 0.56

	precision	recall	f1-score
0	1.00	1.00	1.00
1	1.00	1.00	1.00

0.5295631234476423

0.5668521159898918

Deployment

- A detailed pdf for jupyter notebook has been created and saved containing all the topics mentioned in the slide with details.

Thank You