MIE 1624H Introduction to Data Science and Data Analytics

ASSIGNMENT 2 – SALARY PREDICTION PROBLEM

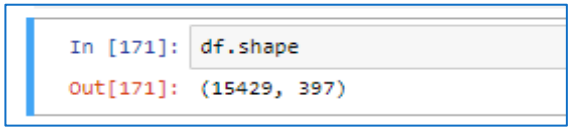
Joshi Urmi Alkesh
(1004822766)

List of Essential Libraries:

- Pandas
- Matplotlib.pyplot
- Numpy
- Sklearn
  - Preprocessing for LabelEncoder
  - Feature_selection for SelectFromModel
  - SVM for LinearSVC
  - Model_Selection
    - Train_test_split
    - GridSearchCV
  - Linear_Model
    - Lasso
    - Ridge
  - Ensemble
    - GradientBoostingRegressor
    - RandomForestRegressor
- Seaborn

Data Set used:

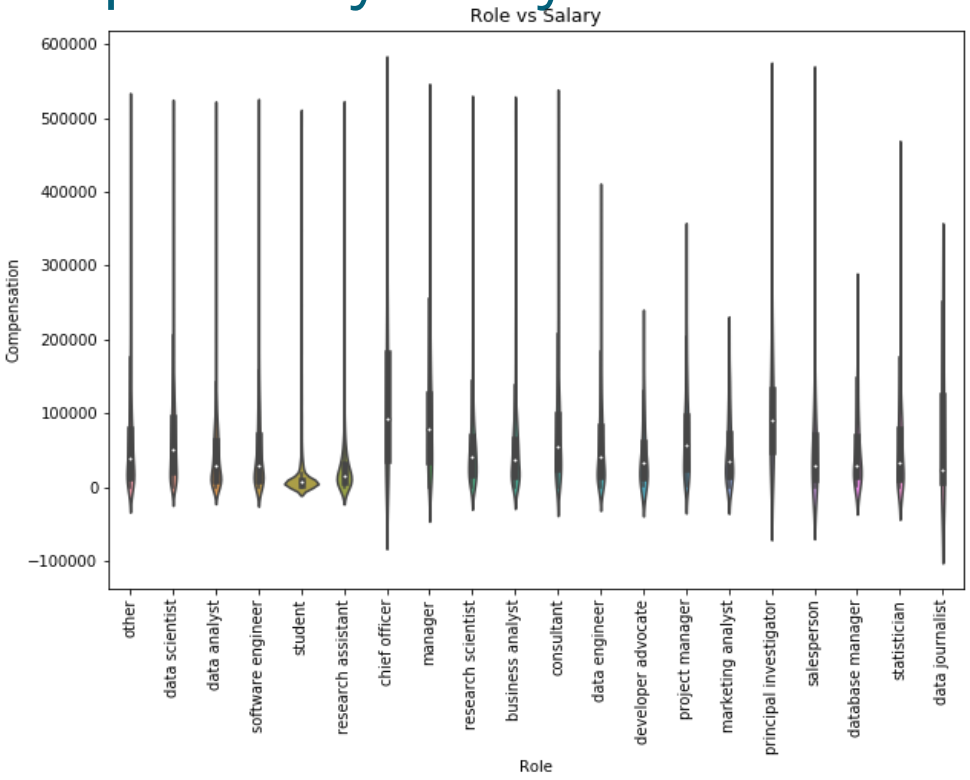
- Kaggle_Salary Data
- Number of Observations:

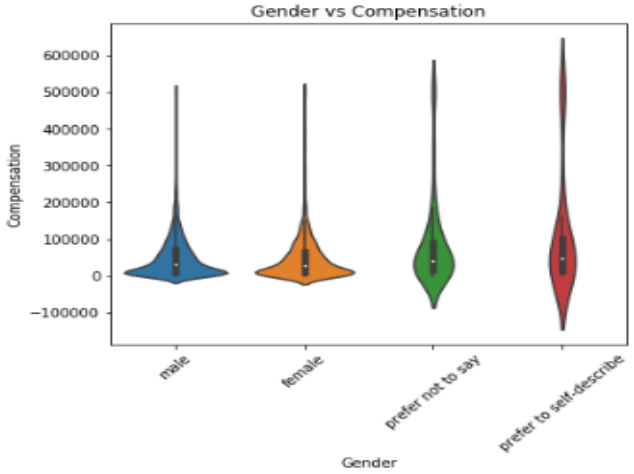


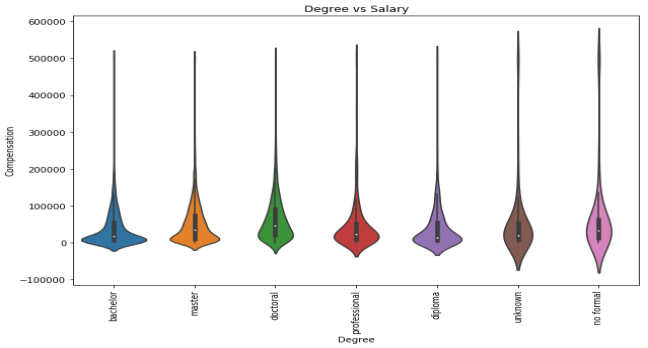
Features used:

- Gender, Age, Country, Degree, Field, Role, Industry, Experience, Software (used), IDE, Programming Language, Recommended Programming Language, Machine Learning Algorithms, Data Visualization, Coding Time, Coding Experience, ML Experience, Data Type, Expertise , Unfair Bias, Insights.
- Compensation is the target Variable

Exploratory Analysis





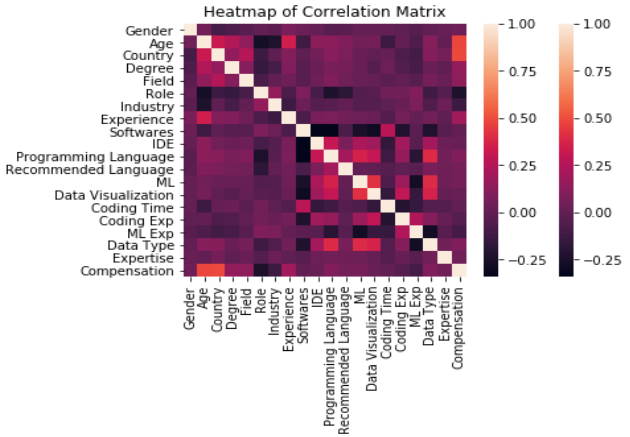


Gender vs Compensation Analysis : explains the influence of gender that takes place on compensation. We can say that male have more influence.

Degree vs Compensation Analysis : explains influence of role on salary. With increase in degree (Bachelor's to Doctoral) difference is also observed in Salary.

Role vs Compensation : defines role influence on compensation.

Correlation Matrix : It gives relationship statistics between all the features. But here as most of the features are categorical, therefore heatmap will not show relation.



Categorical To Numerical Data

- Categorical variables are known to hide and mask lots of interesting information in a data set.
- There are some algorithms which do not convert categorical data automatically.
- Therefore, explicitly it is converted.
- Here, LabelEncoder is used. It is one of the sklearn model. It is easier to use when dealing with small set of features.
- One of its cons is, Labels are dependent to each other i.e. Weights are assigned to each category.

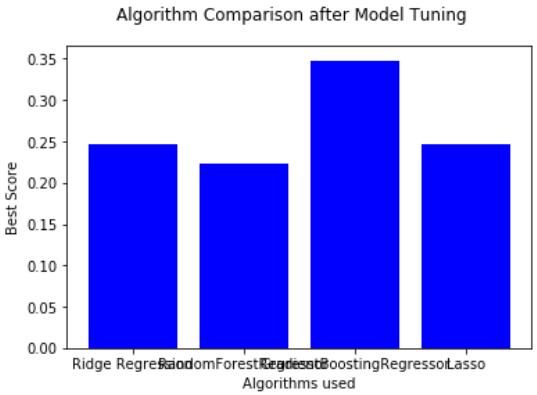
Model Feature Importance

- Second important step here is. Feature selection.
- As, there are many features all are related to each other. For example, Q16 parts and Q17 are related.
- Many of the features are not important but still present.
- False feature selection can weak up the model, else will make the model more accurate.

Model Selection

- Model selection is an important part of any statistical analysis, and indeed is central to the pursuit of science in general.
- After testing 4 models selected,
  - Ridge
  - Lasso
  - HuberRegressor
  - GradientBoostingRegressor

It is found that Gradient Boosting shows more accuracy.



Model Results:

Before Model Tuning

- It is observed that ridge, lasso gives some what similar over-fitting. While GradientBoosting gives maximum over-fitting.
- GradientBoosting is found to be the best model

Model	Training set accuracy (%)	Testing set accuracy (%)	Variance
Ridge	24.23	28.56	0.0013
Lasso	24.24	28.56	0.0014
Huber Regressor	21.43	25.17	0.0012
Gradient Boosting Regressor	32.94	38.30	0.0020

After Model Tuning

- It is observed that ridge, lasso gives some what similar over-fitting. While GradientBoosting gives maximum over-fitting.
- Therefore, it shows similar as that model implementation. Just the accuracy gets improved.
- GradientBoosting is found to be the best model

performing these steps following are the results obtained:

Model	Training set accuracy (%)	Testing set accuracy (%)	Optimized Training set accuracy (%)	Optimized Training set accuracy (%)
Ridge	24.23	28.56	24.58	28.54
Lasso	24.23	28.56	24.59	25.84
Huber Regressor	21.43	25.17	22.23	24.97
Gradient Boosting Regressor	32.94	28.30	34.79	39.83