Data Jobs Salaries

Merve Gürbüz
Egecan Serbester

Feature Analysis with HeatMap

PCA for Dimension Reduction

- Labelled the String columns (Used LabelEncoder)
- Experiment the PCA (4 - 6 - 7 - 8 - 10)

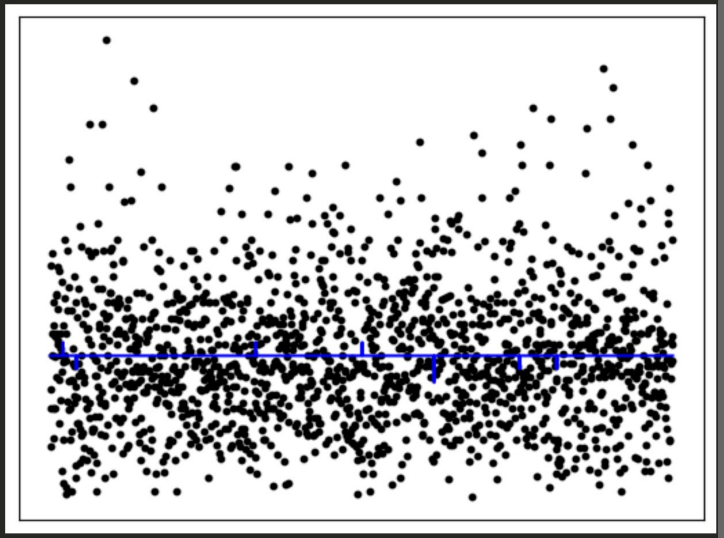
Dimension	MSE for Standard Scale Data
4	0.7418001636406454
6	0.5774320163069093
7	0.11008455543984688
8	0.14195711453268917
10 (real dimension)	0.12366654622069748

Row Labels	Count of job_title
CT	18
FL	11
FT	7932
PT	13
Grand Total	**7974**

Row Labels	Count of job_title
2020	75
2021	218
2022	1650
2023	6031
Grand Total	**7974**

Row Labels	Count of job_title
EN	429
EX	245
MI	1546
SE	5754
Grand Total	**7974**

```
employment_type_numeric
Coefficients:
 [[12783.99921654]]
Mean squared error: 4240747051.33
Coefficient of determination: -0.00
```



- Data is imbalance!
- Linear regression is meaningless since there isn't any column that directly correlate to the salary.

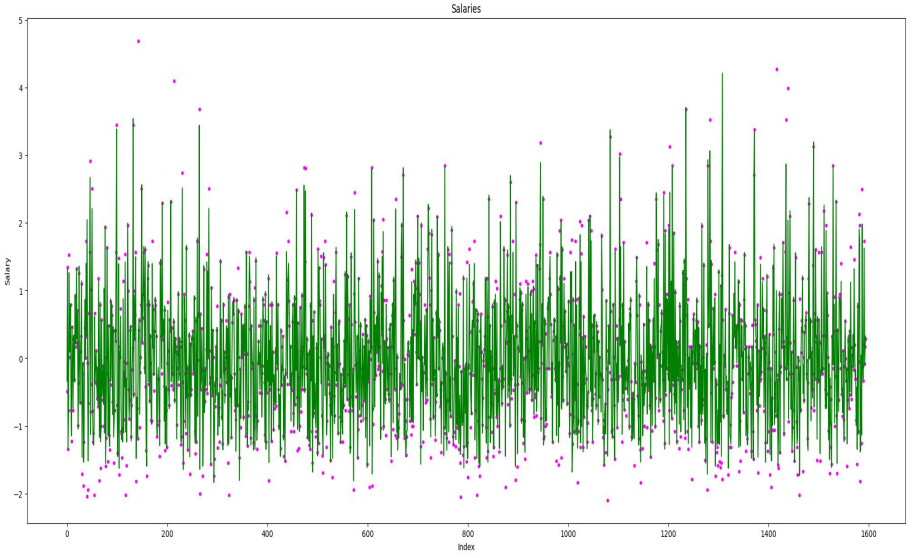
SVR

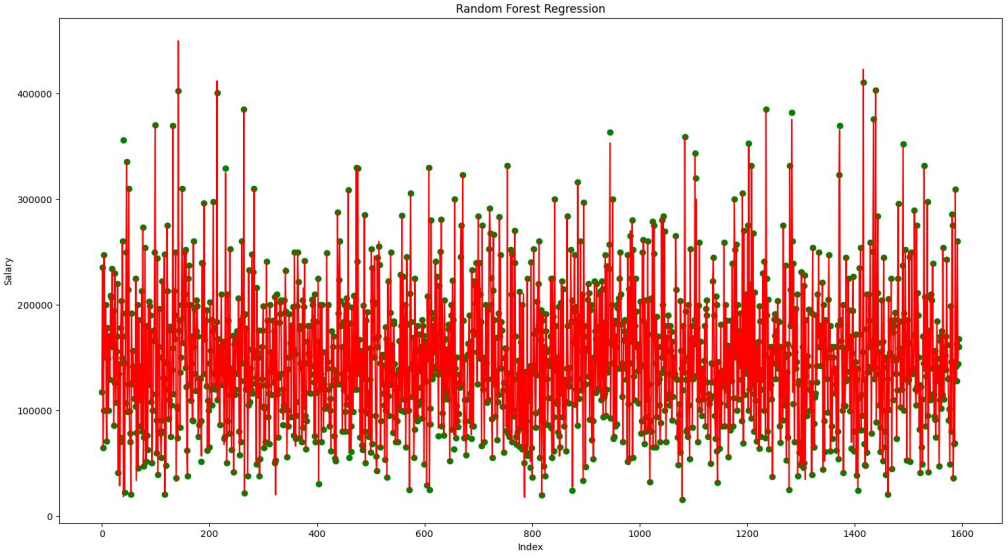
- Best Kernel: RBF (Radial Basis Function)
- RMSE : 65342
- Standardized RMSE: 0.36

Random Forest Regression

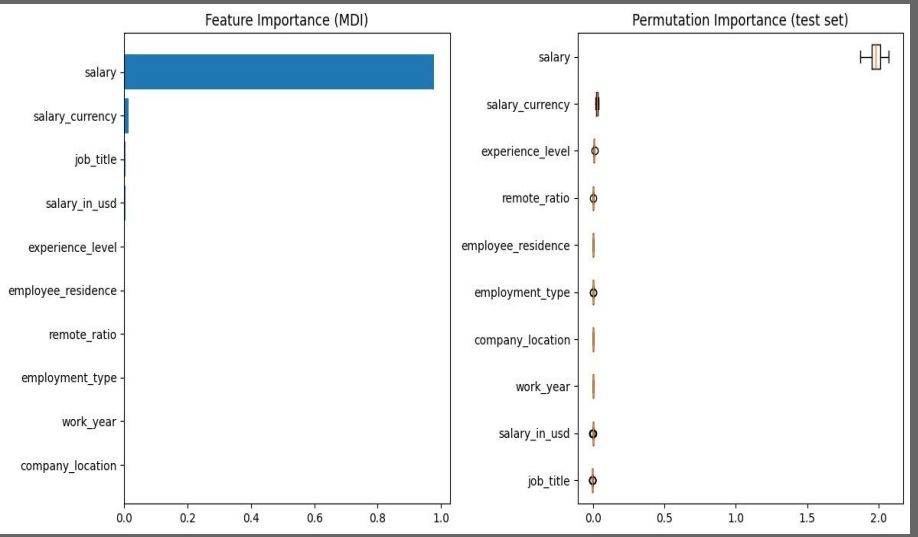
Tried different parameters with GridSearchCV(regr, parameters)
'n estimators': [100, 150, 200, 250, 300], 'max_depth':[1,2,3,4] (best for n=300, d=4)

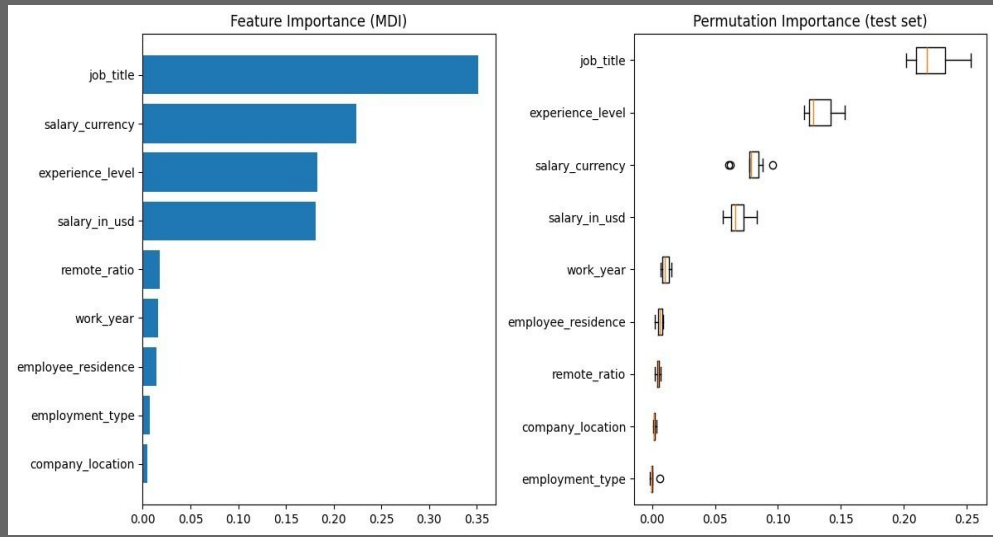
N	Standardized RMSE	RMSE
7	0.1572816	10128
10	0.161137	10117





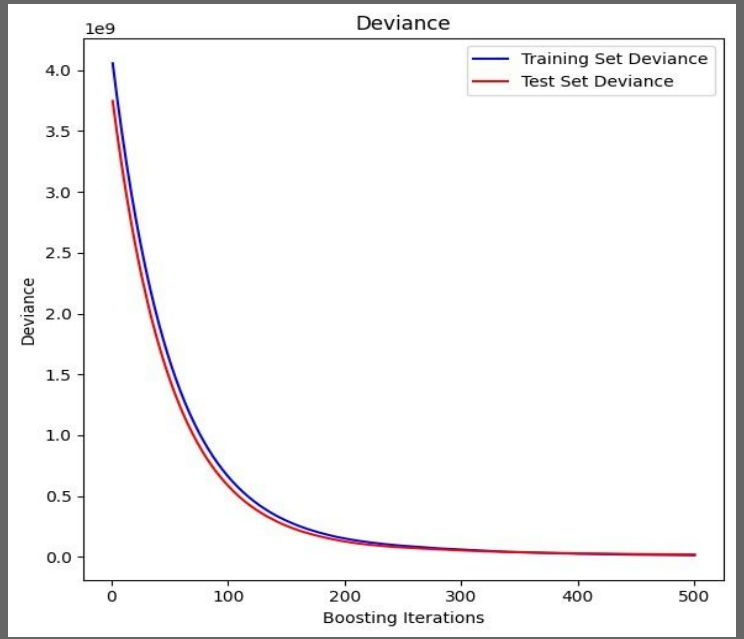
Feature Importance





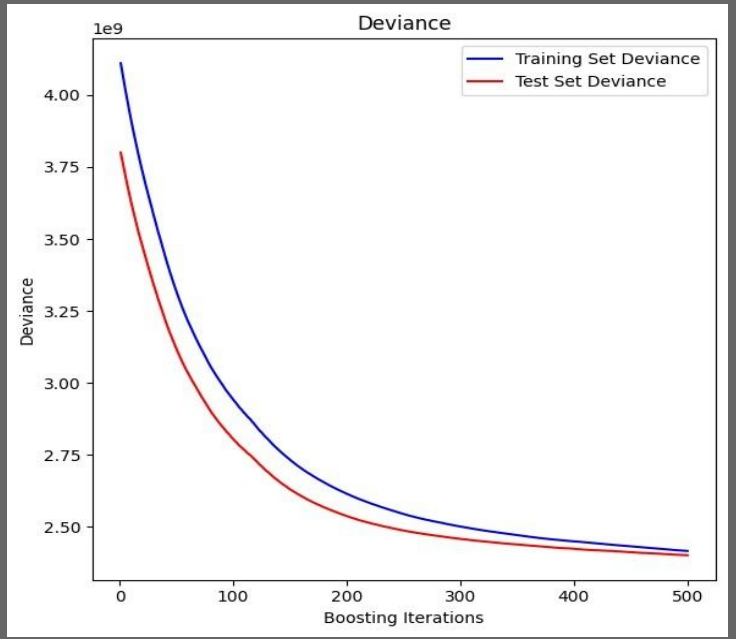
XG Boost

Training and Test Sets Overfit



RMSE: 48957

more reasonable without salary



RMSE: 4573

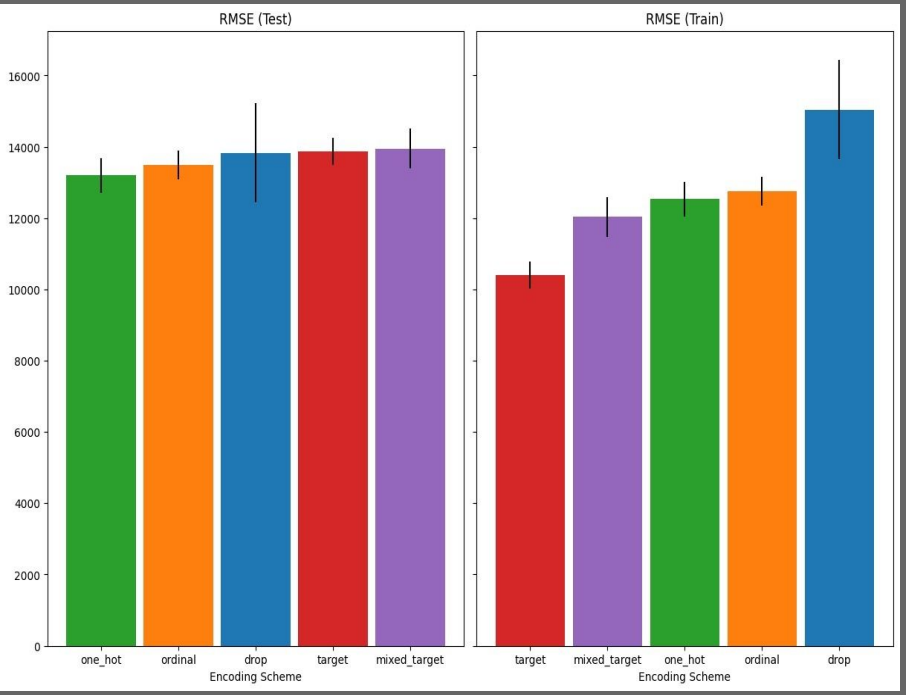
Ordinary Least Square Regression

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.669e+07	2.43e+06	-10.974	0.000	-3.15e+07	-2.19e+07
work_year	1.32e+04	1203.117	10.975	0.000	1.08e+04	1.56e+04
experience_level_numeric	1.669e+04	835.339	19.980	0.000	1.51e+04	1.83e+04
employment_type_numeric	5404.0621	5855.759	0.923	0.356	-6074.759	1.69e+04
job_title_numeric	492.2807	25.278	19.475	0.000	442.729	541.832
salary	0.0077	0.001	5.628	0.000	0.005	0.010
employee_residence_numeric	442.2196	122.894	3.598	0.000	201.315	683.124
remote_ratio	-37.8901	13.598	-2.787	0.005	-64.545	-11.235
company_location_numeric	473.9454	140.631	3.370	0.001	198.272	749.619
company_size_numeric	-7189.9150	2018.388	-3.562	0.000	-1.11e+04	-3233.346

```python
#prediction function with respect to results above
def calculatePredictedSalary(work_year,experience_level_numeric,employment_type_numeric,job_title_numeric,
                             salary,employee_residence_numeric,remote_ratio,company_location_numeric,company_size_numeric):
    prediction = -2.669e+07 + work_year*(1.32e+04) + experience_level_numeric*(1.669e+04)
    prediction += 5404.0621*employment_type_numeric  + job_title_numeric*492.2807 + salary*0.0077
    prediction += 442.2196*employee_residence_numeric + (-37.8901)*remote_ratio + company_location_numeric*473.9454 -7189.9150*company_size_numeric
    return prediction
#but because coef is -2.669e+07, this data is not suitable for that prediction
```

Selecting The Best Encoder

Hist Gradient Boosting Regressor



GBoost