CRIME DATA ANALYSIS USING REGRESSION ALGORITHMS

SIDDHARTH MANDGI

ABHISHEK LOKAM

WHAT WE HAVE DONE

Taking a crime data set and predicting the victim descent and victim

- We have a data set for crime from 2011 to 2018
- From this data set we extracted data for 2017 and 2018
- Then we cleaned the data set
- Prediction of Victim Descent and Victim Sex

OPTIMIZATION MODEL

- We took four algorithms namely:

1) Linear Regression

2) Decision Tree

3) Random Forest

4) Bagging

- Predicted the Victim Descent using these four algorithms

FLOW OF OUR PYTHON CODE

IMPORTING DATA

- Importing Crime Data Set for all Years from 2011 -2018

```
In [89]: df = pd.read_csv('Crimedata.csv')
         df
```

Out[89]:

	DR Number	Date Reported	Date Occurred	Time Occurred	Area ID	Area Name	Reporting District	Crime Code	Crime Code Description	MO Codes	...	Weapon Description	Status Code	St Descrip
0	1208575	03/14/2013	03/11/2013	1800	12	77th Street	1241	626	INTIMATE PARTNER - SIMPLE ASSAULT	0416 0446 1243 2000	...	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	AO	Adult I
1	102005556	01/25/2010	01/22/2010	2300	20	Olympic	2071	510	VEHICLE - STOLEN	NaN	...	NaN	IC	Invest
2	418	03/19/2013	03/18/2013	2030	18	Southeast	1823	510	VEHICLE - STOLEN	NaN	...	NaN	IC	Invest
3	101822289	11/11/2010	11/10/2010	1800	18	Southeast	1803	510	VEHICLE - STOLEN	NaN	...	NaN	IC	Invest
4	42104479	01/11/2014	01/04/2014	2300	21	Topanga	2133	745	VANDALISM - MISDEAMEANOR ($399 OR UNDER)	0329	...	NaN	IC	Invest
5	120125367	01/08/2013	01/08/2013	1400	1	Central	111	110	CRIMINAL HOMICIDE	1243 2000 1813 1814 2002 0416 0400	...	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	AA	Adult A
6	101105609	01/28/2010	01/27/2010	2230	11	Northeast	1125	510	VEHICLE - STOLEN	NaN	...	NaN	IC	Invest
7	101620051	11/11/2010	11/07/2010	1600	16	Foothill	1641	510	VEHICLE - STOLEN	NaN	...	NaN	IC	Invest

DATA FILETRING

NEXT WE CLEAN THE DATA SET

	Lat	Lon	Time	Weekday	Area ID	Crime Code	Victim Descent	Victim Age	Victim Sex
1308923	34.0886	-118.2979	19.500000	4	2	510	X	16.0	U
1382661	34.0512	-118.2787	17.000000	5	2	510	X	16.0	U
1383229	34.0328	-118.2915	7.750000	1	3	510	X	16.0	U
1383510	34.0676	-118.2202	0.016667	4	4	510	X	16.0	U
1383605	33.7347	-118.2842	7.500000	5	5	510	X	16.0	U
1383932	34.0762	-118.3441	23.000000	0	7	510	X	16.0	U
1384619	34.0510	-118.2480	17.000000	3	1	510	X	16.0	U
1384681	34.0480	-118.2438	20.500000	1	1	330	B	29.0	M
1384950	34.0210	-118.2123	10.000000	3	4	510	X	16.0	U
1385019	34.0699	-118.2595	3.000000	4	2	510	X	16.0	U
1385100	34.0644	-118.2630	0.016667	6	2	510	X	16.0	U
1385132	34.0225	-118.2156	22.500000	0	4	510	X	16.0	U
1385457	34.0544	-118.2767	22.000000	4	2	510	X	16.0	U
1385780	34.0797	-118.2183	18.000000	6	4	510	X	16.0	U
1385864	34.0230	-118.2793	7.166667	4	3	510	X	16.0	U
1385903	34.0428	-118.2532	22.500000	3	1	510	X	16.0	U

PREDICTION

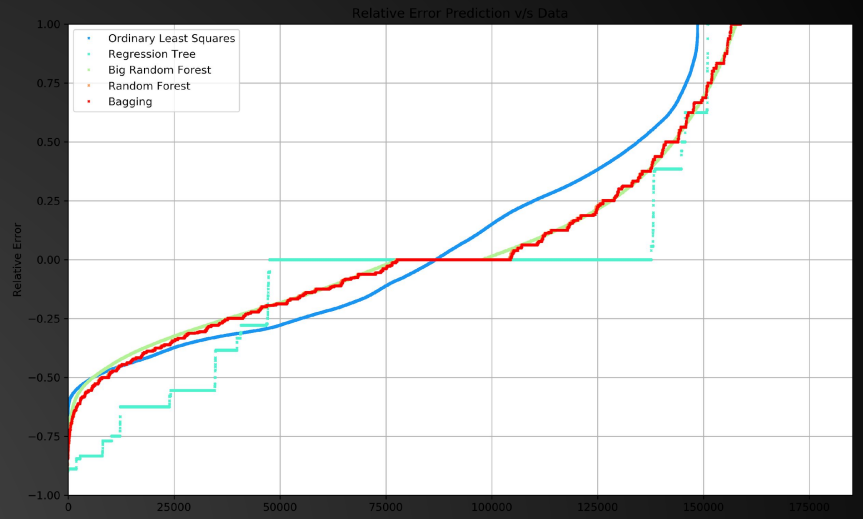
- We then try to predict Victim Sex Or the Victim Descent using four algorithms linear regression , Random forest , Decision Tree and Bagging

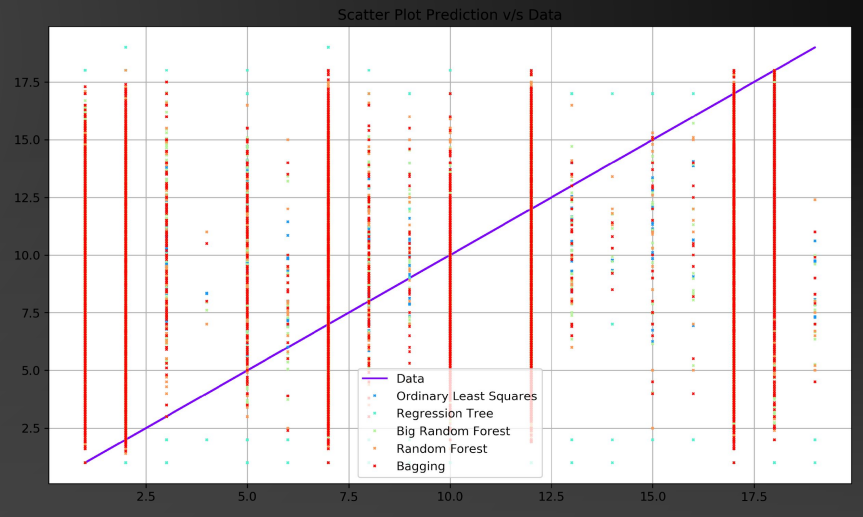
```
for _ in range (4):
    train_X, test_X, train_y, test_y = (
        train_test_split(df2_2017.drop("Victim Sex", axis=1),
                         df2_2017["Victim Sex"], test_size=0.25))
    for i, obj_factory in enumerate(experiments["Objects"]):
        obj = obj_factory()
        obj.fit(y=train_y,X=train_X)
        experiments["Predictions"][i] += list(obj.predict(test_X))
    actuals += list(test_y)
actuals = pd.Series(actuals)
experiments["Predictions"] = list(map(pd.Series, experiments["Predictions"]))
```

RESULTS OF OUR ANALYSIS

RESULTS

VICTIM DESCENT ANALYSIS AND PREDICTION : 2017



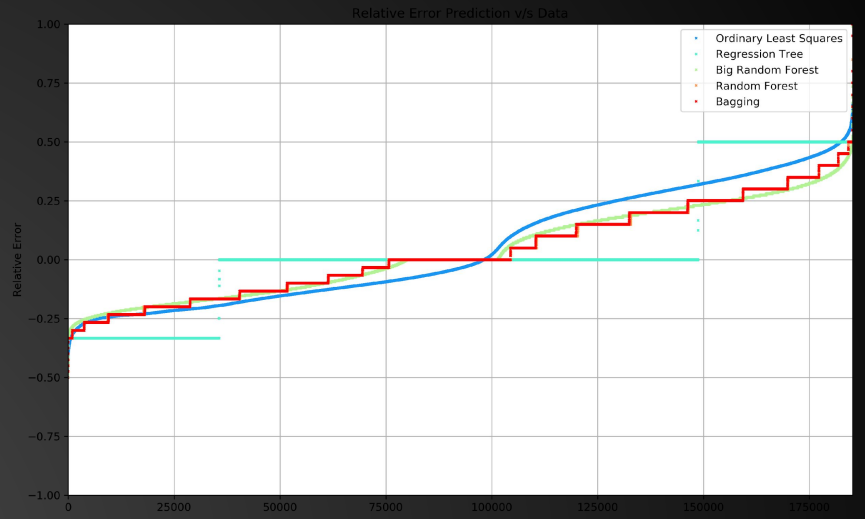


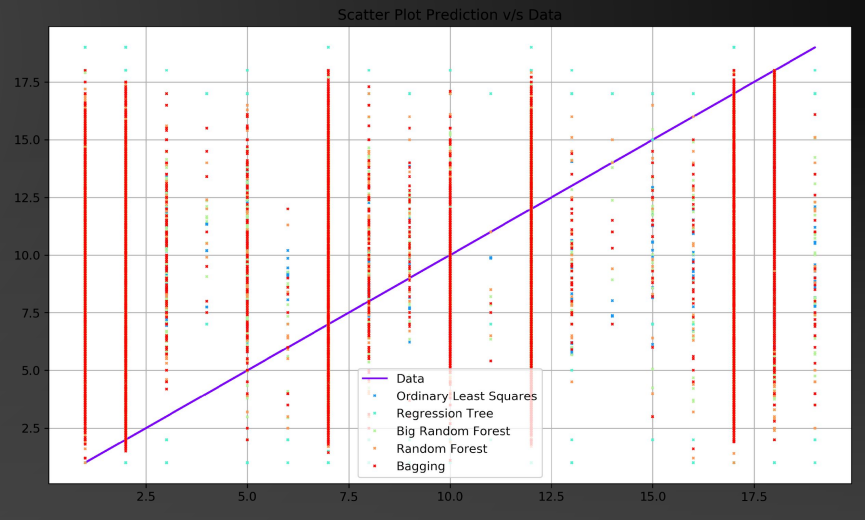
RESULTS

Algorithm	Results
Ordinary Least Squares	0.135275
Regression Tree	-0.137305
Big Random Forest	0.417824
Random Forest	0.366543
Bagging	0.367579

RESULTS

VICTIM DESCENT ANALYSIS AND PREDICTION FOR 2018





RESULTS

Algorithm	Results
Ordinary Least Squares	0.129475
Regression Tree	-0.166361
Big Random Forest	0.400683
Random Forest	0.349482
Bagging	0.348480

CONCLUSION

While comparing all four algorithms while analyzing the crime data set , We find that Big Random Forest yields the best results with the best accuracy

As we increase the estimator values in Random forest algorithm the accuracy increases.

THANK YOU