Wayfair

Cutomer data analysis

Goal of the Project

1. Whether a B2B customer will purchase or not in the next 30 days

2. How much a B2B customer will spend in the next 30 days

Steps to Process

- Data Cleaning
- Data Splitting
- Model Selection
- Model Evaluation
- Feature Importance
- Model Optimization
- Final Output

Data Cleaning

1.First I divided the columns into two parts :

- Categorical Column
- Numerical Column

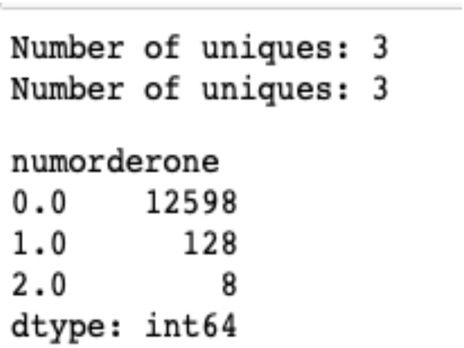
2. 75% of the numerical column has values 0.0

➡ I put 0.0 in null values.

3. The categorical columns are given values 0,1,2,3..to help the operation.
For example:

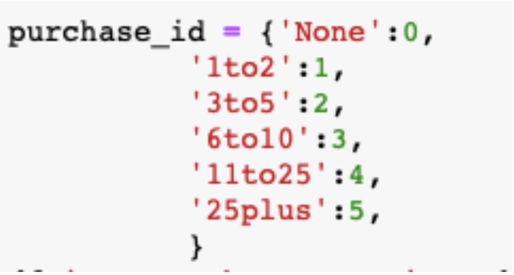
➡ Purchase id : ‘None’ : 0,
’1to2’ : 1,

4.This is how all the columns are turned into numerical columns

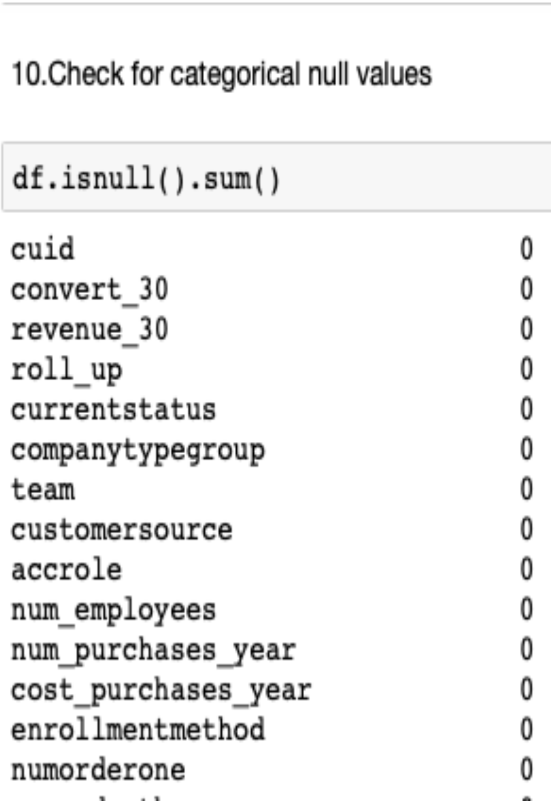


```
Number of uniques: 3
Number of uniques: 3

numorderone
0.0     12598
1.0       128
2.0         8
dtype: int64
```



```
purchase_id = {'None':0,
              '1to2':1,
              '3to5':2,
              '6to10':3,
              '11to25':4,
              '25plus':5,
              }
```



10.Check for categorical null values

```
df.isnull().sum()
```

```
cuid                    0
convert_30              0
revenue_30              0
roll_up                 0
currentstatus           0
companytypegroup        0
team                    0
customersource          0
accrole                 0
num_employees           0
num_purchases_year      0
cost_purchases_year     0
enrollmentmethod        0
numorderone             0
```

Data Splitting

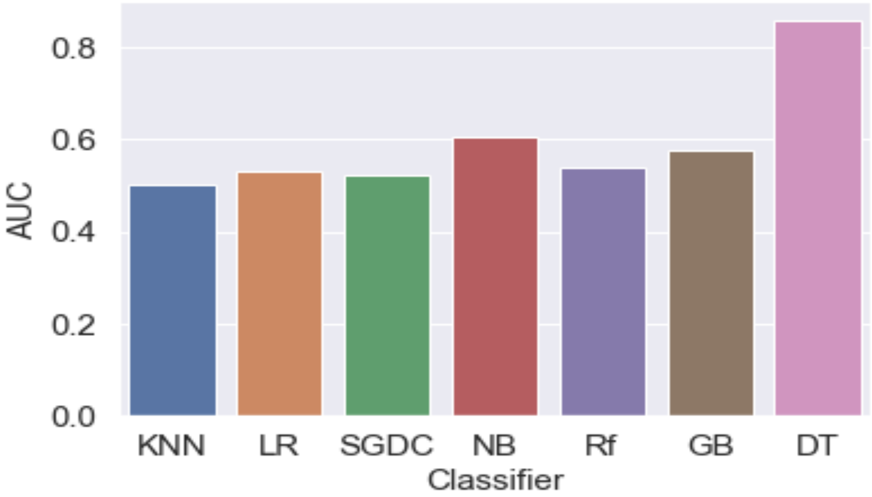
- K fold method
- Train test split

Model Selection

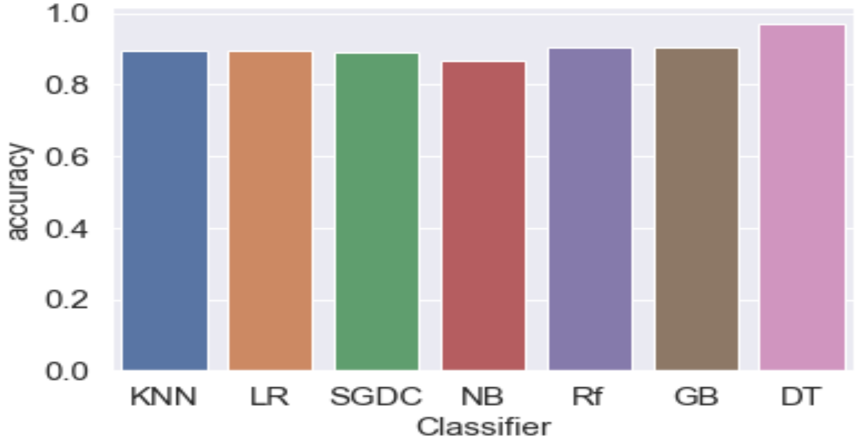
- K Nearest Neighbor (KNN)
- Logistic Regression (LR)
- Stochastic Gradient Descent (SDGC)
- Naïve Byes (NB)
- Random Forest (RF)
- Gradient Boosting (GB)
- Decision Tree (DT)
- Lasso Regression
- ElasticNet Regression
- XGB Boost
- Ridge Regression

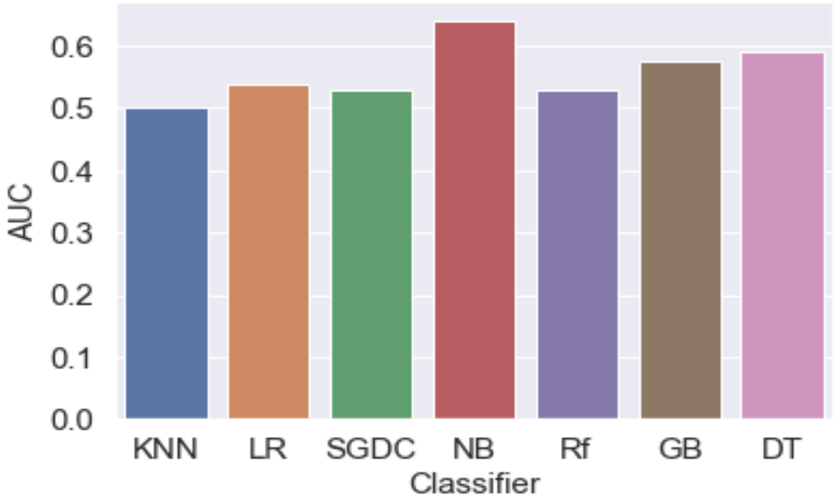
Classification Model Evaluation

For 'Convert_30' based on accuracy and AUC score

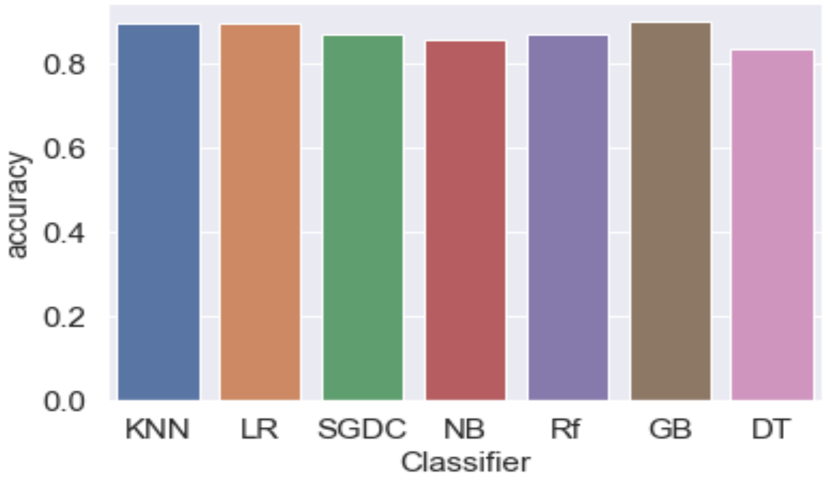


K Fold Method



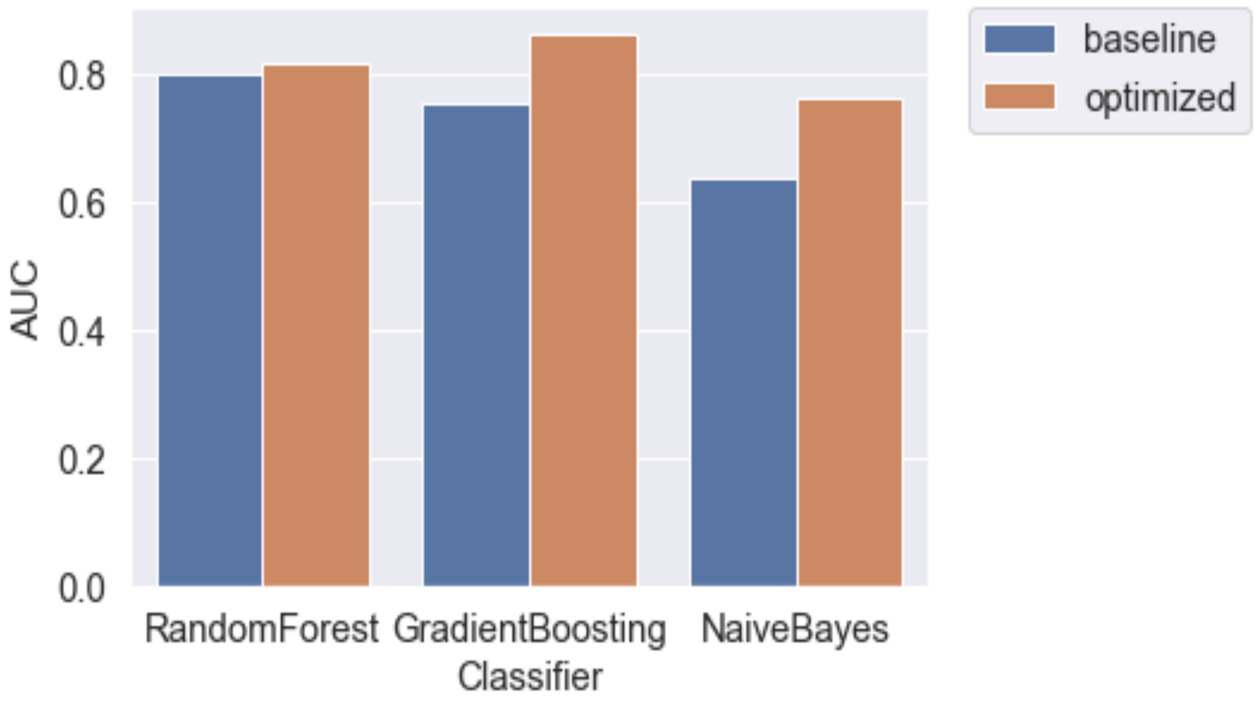


Train Test Split Method



Baseline and Optimization

Classification Problem

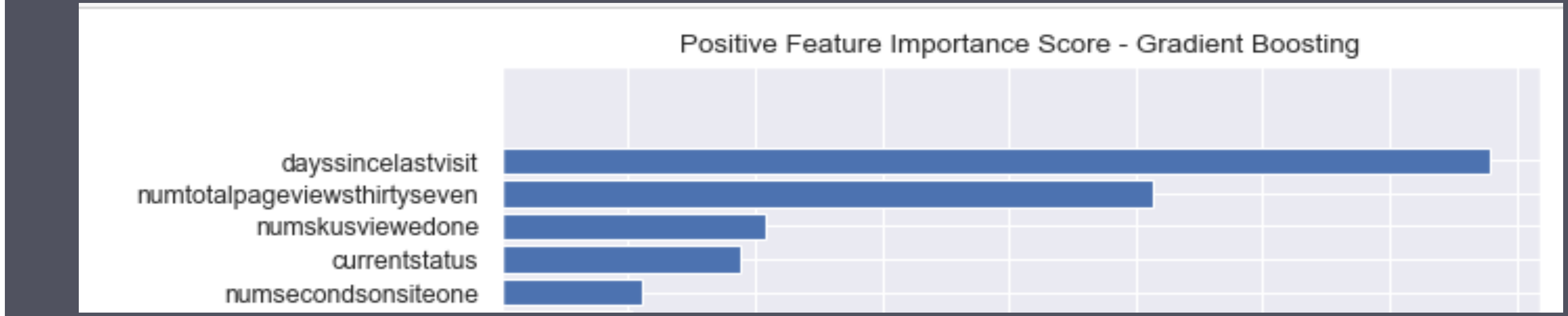


- After analyzing AUC and Accuracy Score the model Random Forest, Gradient Boosting and Naïve Byes are three best models for the classification problem.

- The results shows Gradient Boosting as the best model .

Feature Importance

A Startup PowerPoint Presentation

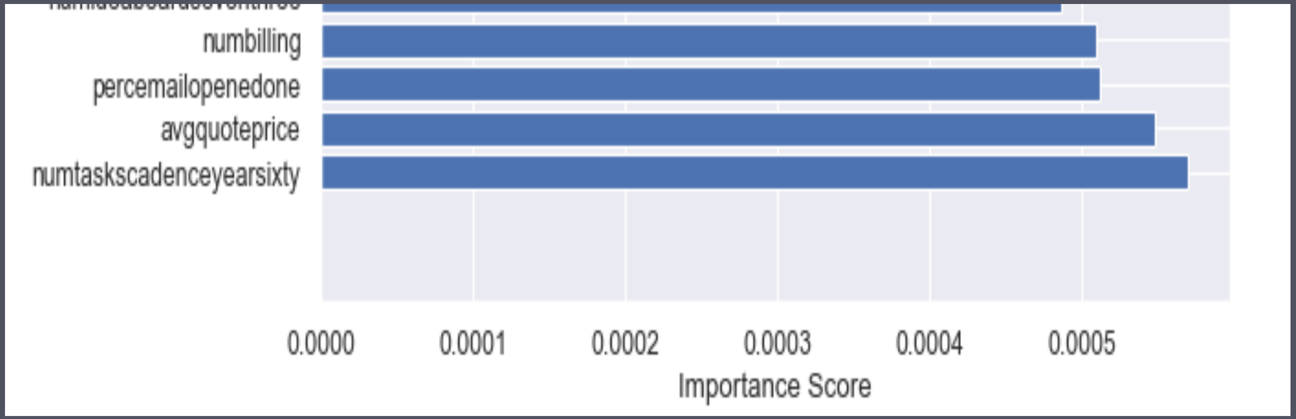


Positively Co-related

- The Top three Positively Correlated features are :
- Days since last visit
- Number of total page view thirty seven days
- Current Status

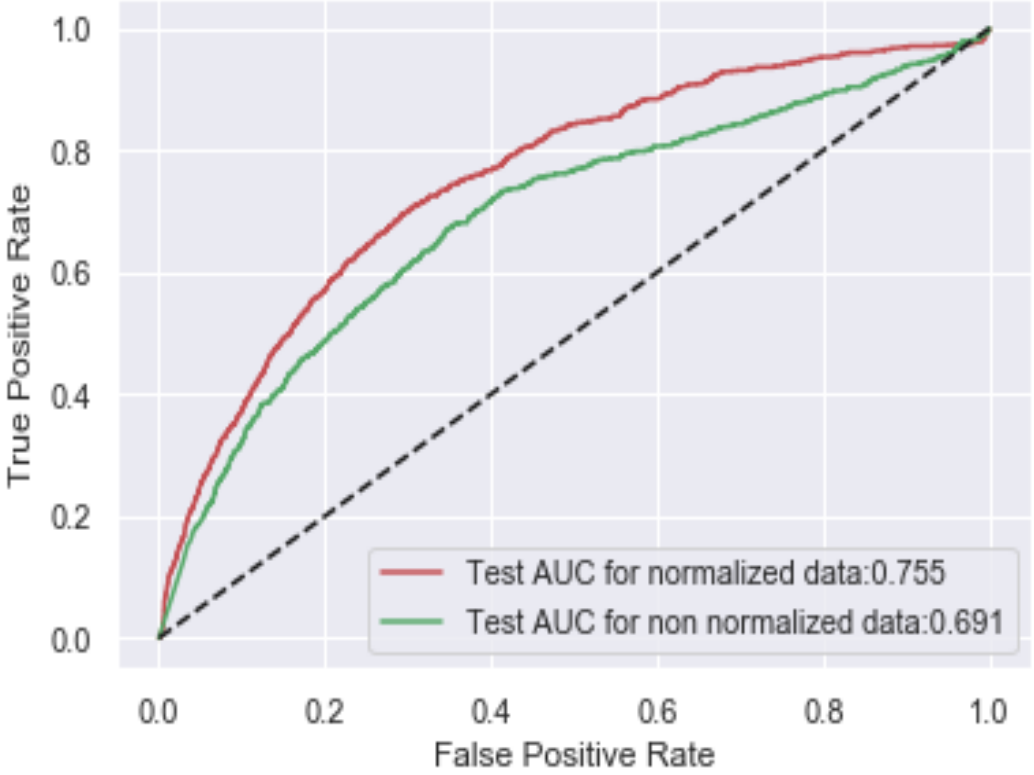
Negatively Co-related

- The Top three Negatively Correlated features are :
- Number of billing
- Average quote price
- Number of tasks attendance in year sixty



AUC ROC Curve

- The Optimized Gradient Boosting Score is 0.891
- According to AUC and ROC curve the normalized data give AUC of 0.755

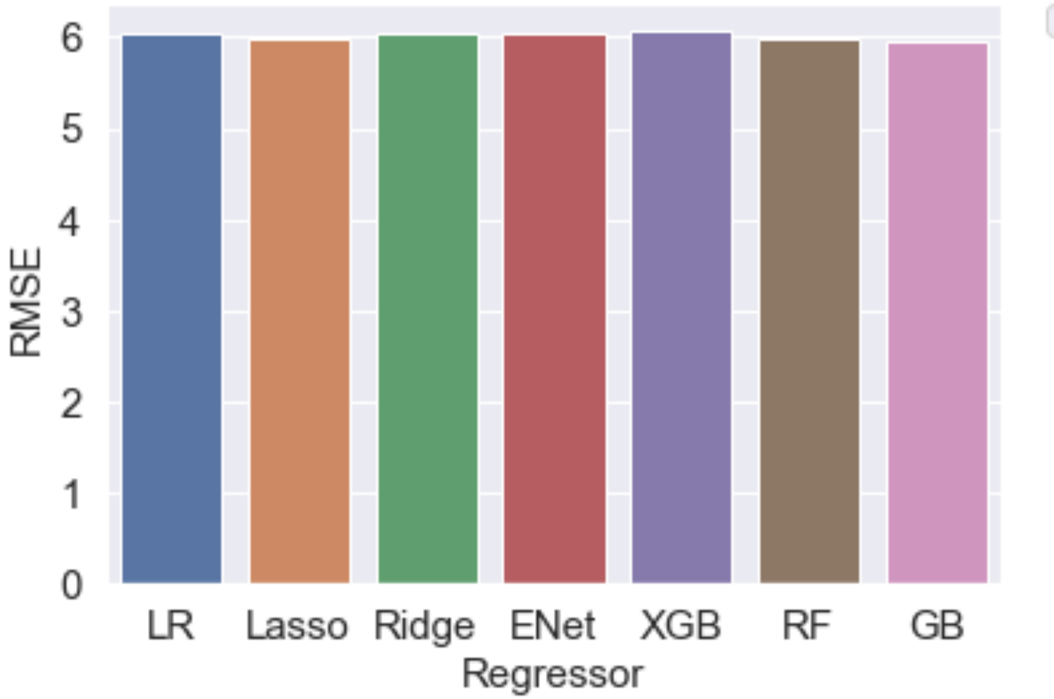


Regression Model Evaluation

Based on Root Means Square Error

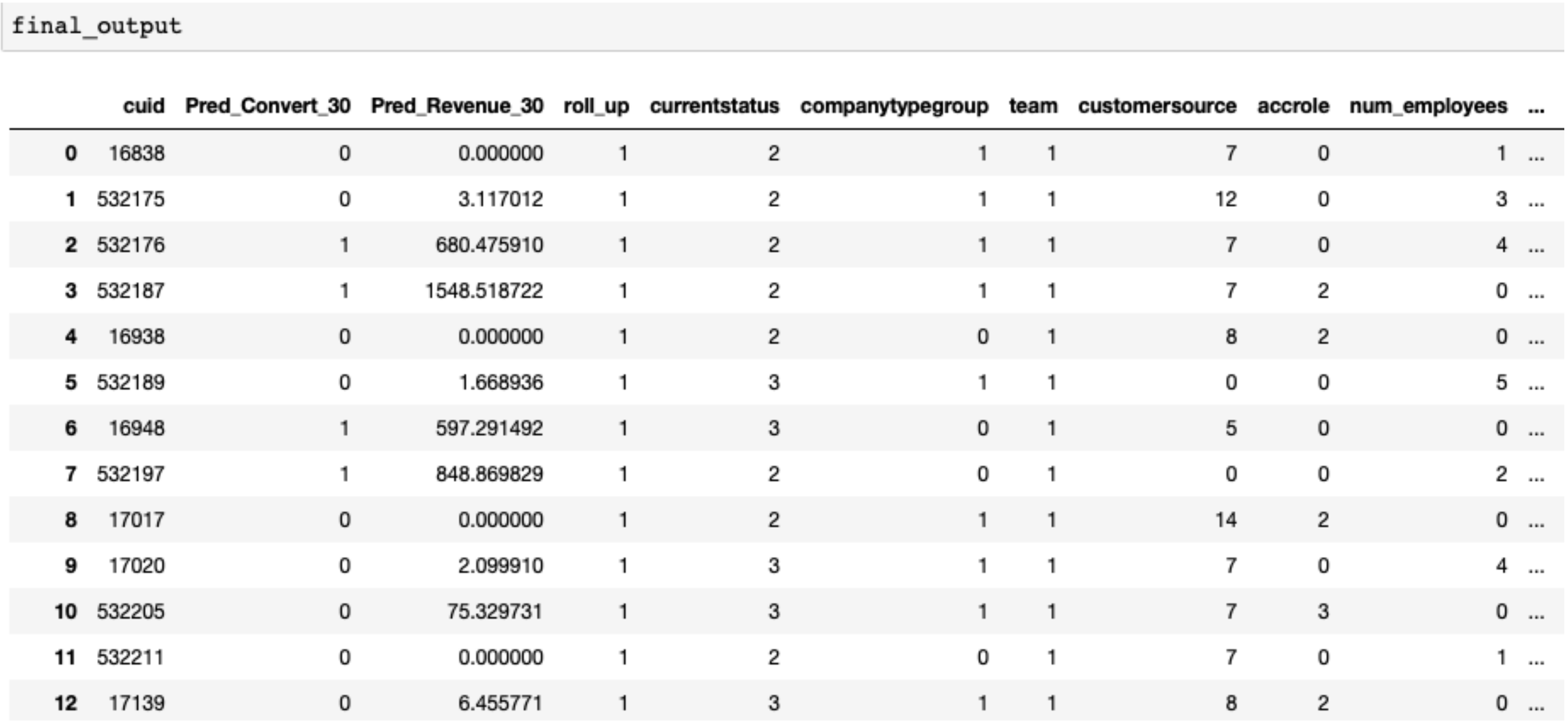
- After Running the dataset through Linear ,Lasso, Ridge, Random Forest and Gradient Boosting egression, we get these rmse scores.

- Gradient Boosting gives lowest rmse score among them



Model Selection and Final Output

For Predicting ‘Convert_30’ and ’Revenue Gradient Boosting algorithm gives out best scores



```
final_output
```

	cuid	Pred_Convert_30	Pred_Revenue_30	roll_up	currentstatus	companytypegroup	team	customersource	accrole	num_employees	...
0	16838	0	0.000000	1	2	1	1	7	0	1	...
1	532175	0	3.117012	1	2	1	1	12	0	3	...
2	532176	1	680.475910	1	2	1	1	7	0	4	...
3	532187	1	1548.518722	1	2	1	1	7	2	0	...
4	16938	0	0.000000	1	2	0	1	8	2	0	...
5	532189	0	1.668936	1	3	1	1	0	0	5	...
6	16948	1	597.291492	1	3	0	1	5	0	0	...
7	532197	1	848.869829	1	2	0	1	0	0	2	...
8	17017	0	0.000000	1	2	1	1	14	2	0	...
9	17020	0	2.099910	1	3	1	1	7	0	4	...
10	532205	0	75.329731	1	3	1	1	7	3	0	...
11	532211	0	0.000000	1	2	0	1	7	0	1	...
12	17139	0	6.455771	1	3	1	1	8	2	0	...