Supervised Learning Capstone Presentation

IBM Employee Attrition.

It's always hard to find the perfect employee. But to keep him happy and satisfied it's a harder task. Any company invests so much time and money to hire, to teach and to keep an employee. Therefore turn to our predictive modeling capabilities and see if we can predict employee attrition on this synthetically generated IBM dataset.

And on the other hand, knowing what factors can possibly to make you unhappy with your job will help you with finding another one.

DataSet

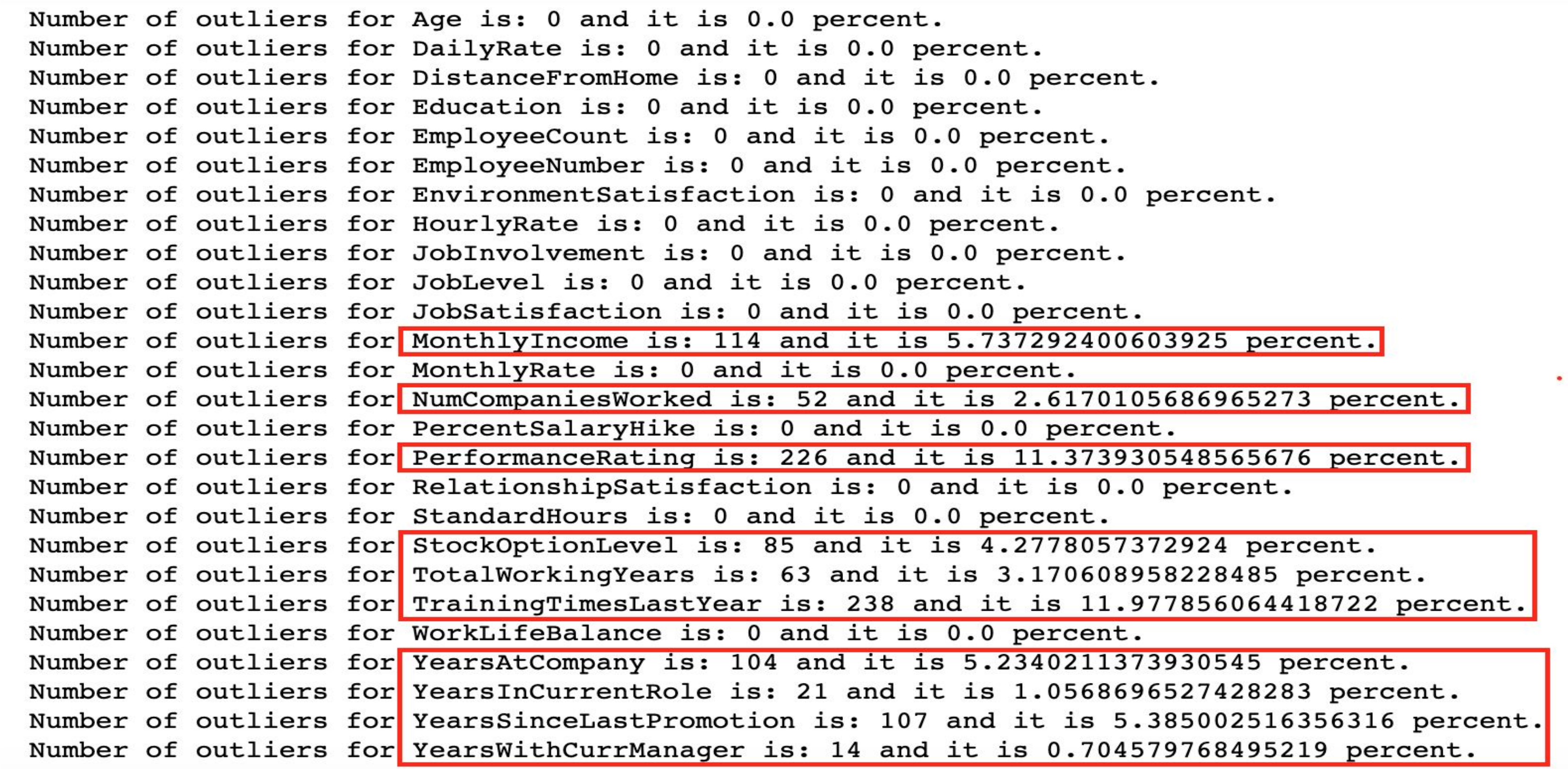
Rows: 1470

Data columns: 35

dtypes: int64(26), object(9)

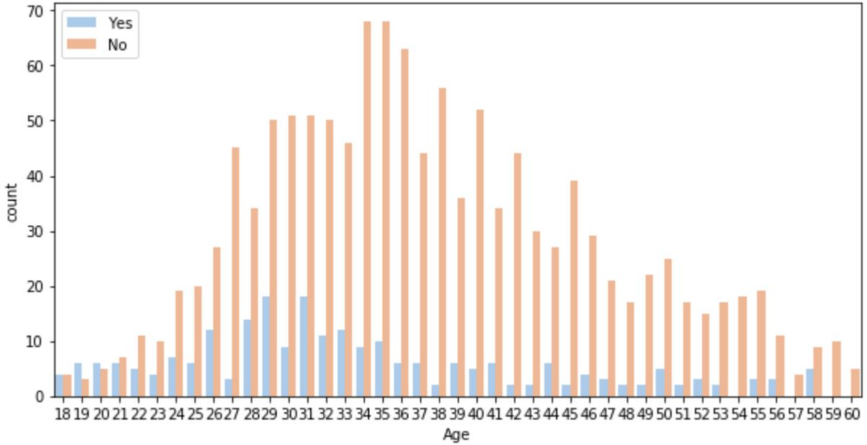
Because of this data set is a fictional created by IBM data scientists, it doesn’t have any missing values. So I can say I got lucky I don’t have to deal with missing values and just move further.

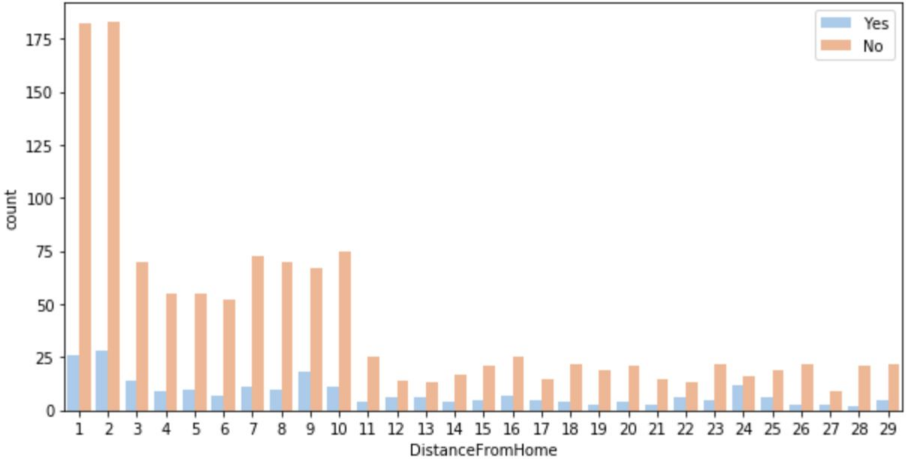
Outliers

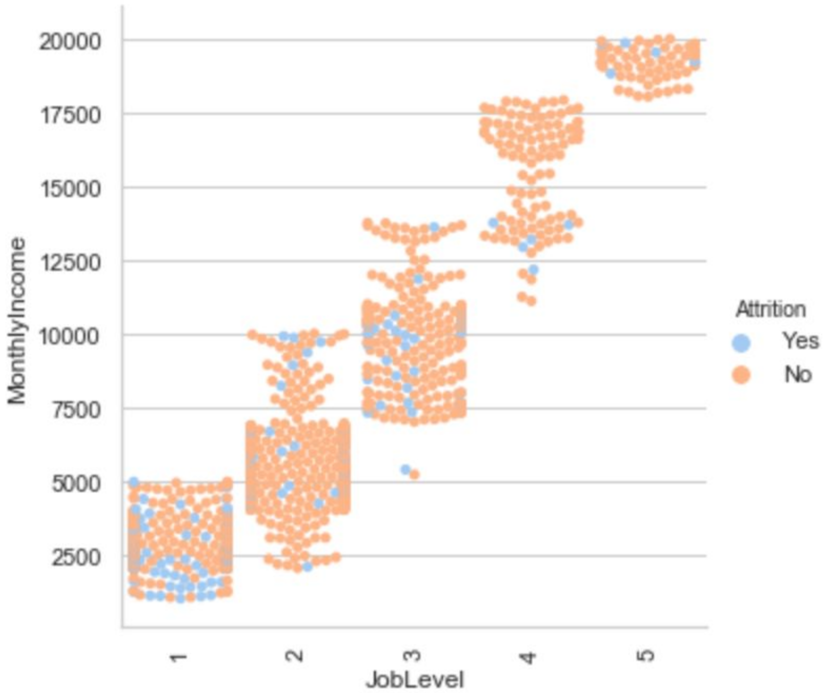


```
Number of outliers for Age is: 0 and it is 0.0 percent.
Number of outliers for DailyRate is: 0 and it is 0.0 percent.
Number of outliers for DistanceFromHome is: 0 and it is 0.0 percent.
Number of outliers for Education is: 0 and it is 0.0 percent.
Number of outliers for EmployeeCount is: 0 and it is 0.0 percent.
Number of outliers for EmployeeNumber is: 0 and it is 0.0 percent.
Number of outliers for EnvironmentSatisfaction is: 0 and it is 0.0 percent.
Number of outliers for HourlyRate is: 0 and it is 0.0 percent.
Number of outliers for JobInvolvement is: 0 and it is 0.0 percent.
Number of outliers for JobLevel is: 0 and it is 0.0 percent.
Number of outliers for JobSatisfaction is: 0 and it is 0.0 percent.
Number of outliers for MonthlyIncome is: 114 and it is 5.737292400603925 percent.
Number of outliers for MonthlyRate is: 0 and it is 0.0 percent.
Number of outliers for NumCompaniesWorked is: 52 and it is 2.6170105686965273 percent.
Number of outliers for PercentSalaryHike is: 0 and it is 0.0 percent.
Number of outliers for PerformanceRating is: 226 and it is 11.373930548565676 percent.
Number of outliers for RelationshipSatisfaction is: 0 and it is 0.0 percent.
Number of outliers for StandardHours is: 0 and it is 0.0 percent.
Number of outliers for StockOptionLevel is: 85 and it is 4.2778057372924 percent.
Number of outliers for TotalWorkingYears is: 63 and it is 3.170608958228485 percent.
Number of outliers for TrainingTimesLastYear is: 238 and it is 11.977856064418722 percent.
Number of outliers for WorkLifeBalance is: 0 and it is 0.0 percent.
Number of outliers for YearsAtCompany is: 104 and it is 5.2340211373930545 percent.
Number of outliers for YearsInCurrentRole is: 21 and it is 1.0568696527428283 percent.
Number of outliers for YearsSinceLastPromotion is: 107 and it is 5.38500251635316 percent.
Number of outliers for YearsWithCurrManager is: 14 and it is 0.704579768495219 percent.
```

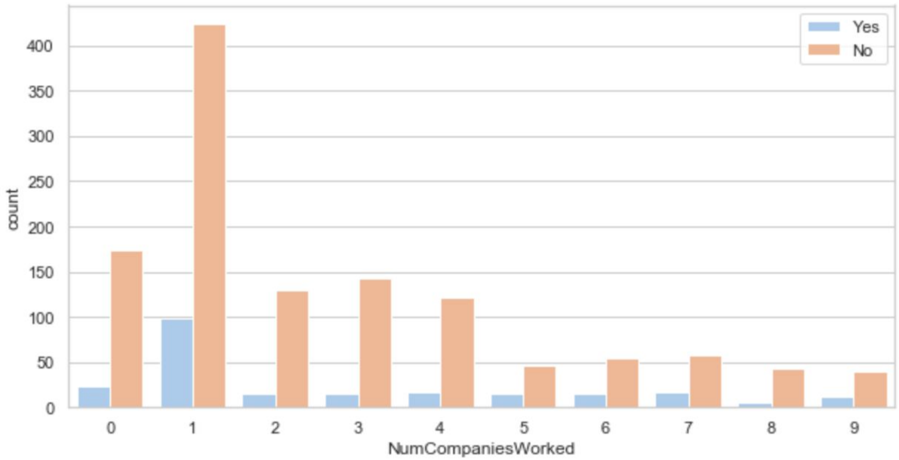
EDA


Yes
No
count
0
10
20
30
40
50
60
70
18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
Age

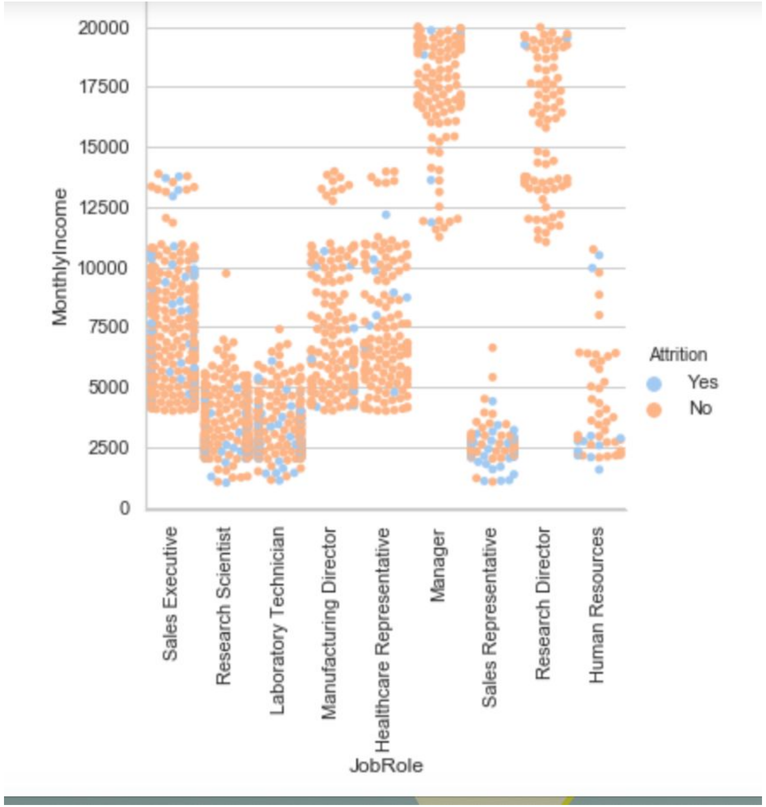

Yes
No
count
0
25
50
75
100
125
150
175
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
DistanceFromHome

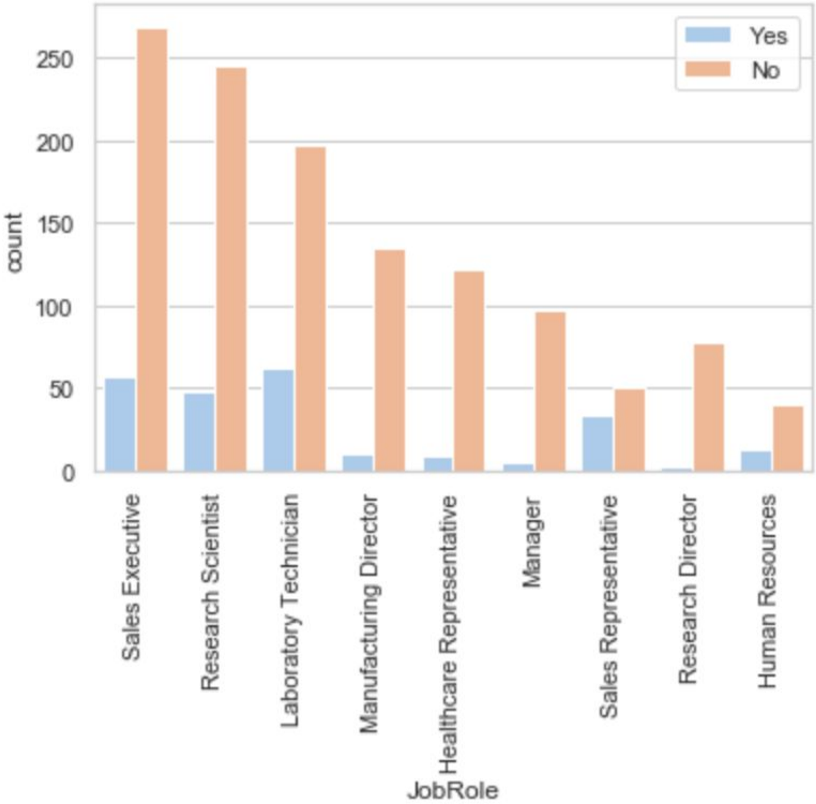


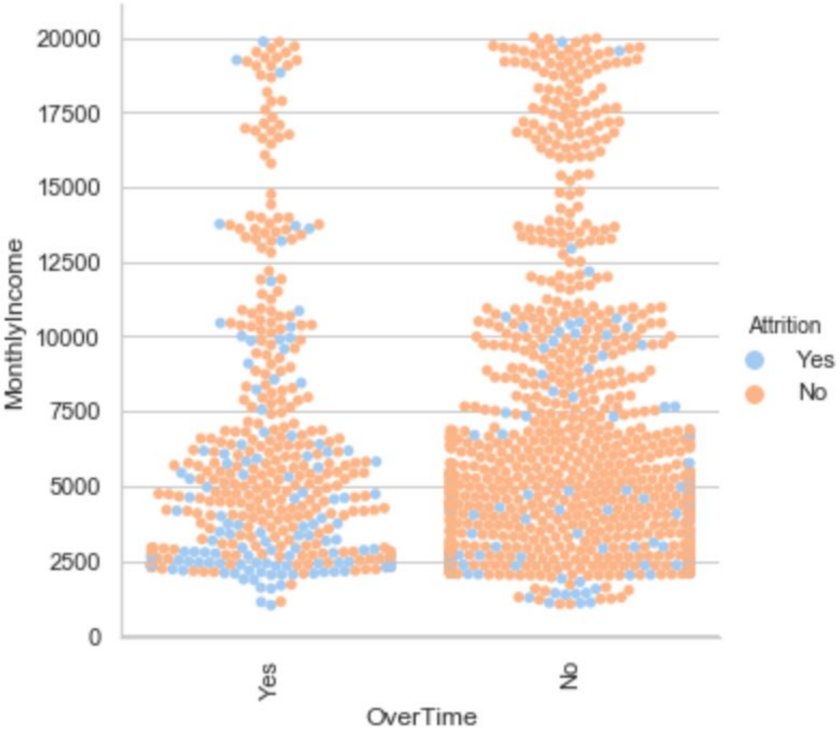
MonthlyIncome
20000
17500
15000
12500
10000
7500
5000
2500
1
2
3
4
5
JobLevel
Attrition
Yes
No

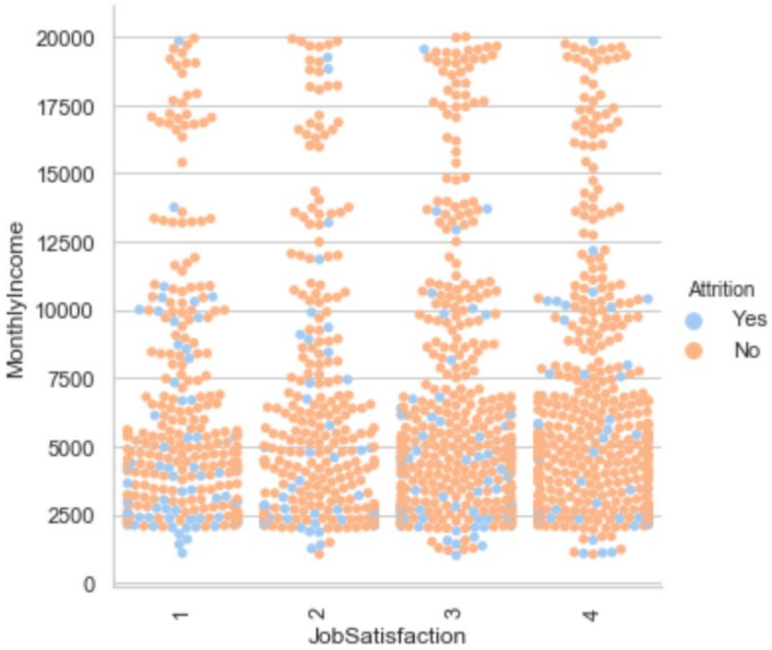


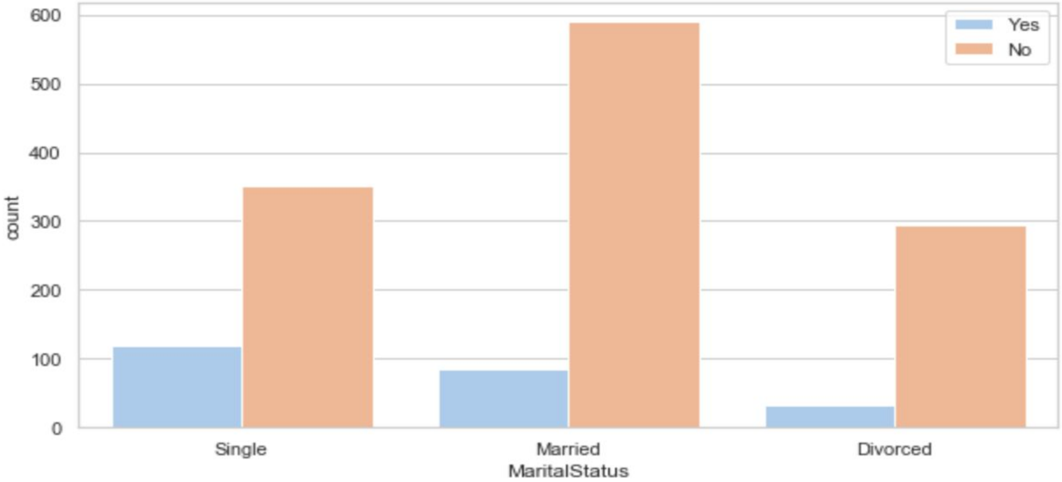
count
400
350
300
250
200
150
100
50
0
Yes
No
0
1
2
3
4
5
6
7
8
9
NumCompaniesWorked

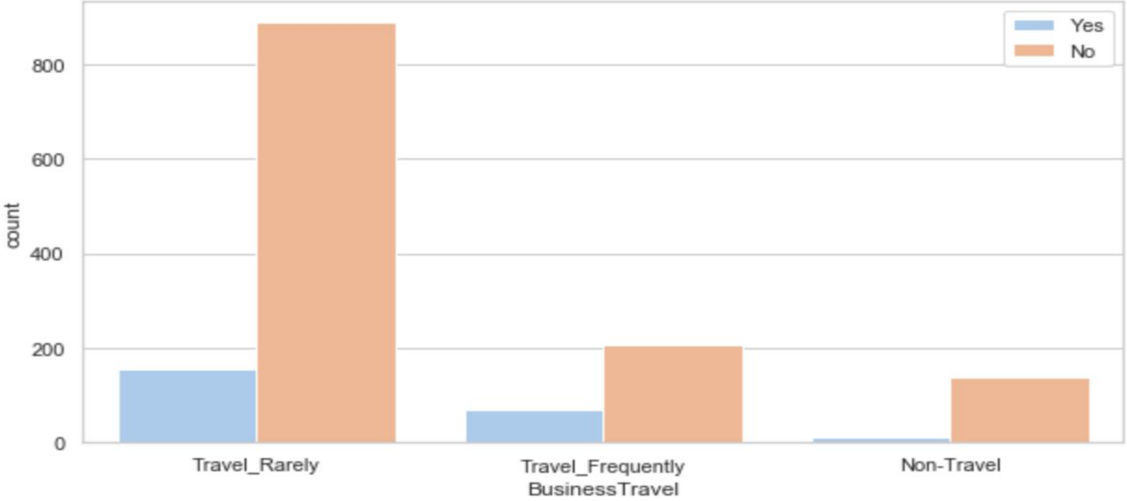

20000
17500
15000
12500
10000
7500
5000
2500
0
MonthlyIncome
Sales Executive
Research Scientist
Laboratory Technician
Manufacturing Director
Healthcare Representative
Manager
Sales Representative
Research Director
Human Resources
JobRole
Attrition
Yes
No


Yes
No
250
200
150
100
50
0
count
Sales Executive
Research Scientist
Laboratory Technician
Manufacturing Director
Healthcare Representative
Manager
Sales Representative
Research Director
Human Resources
JobRole

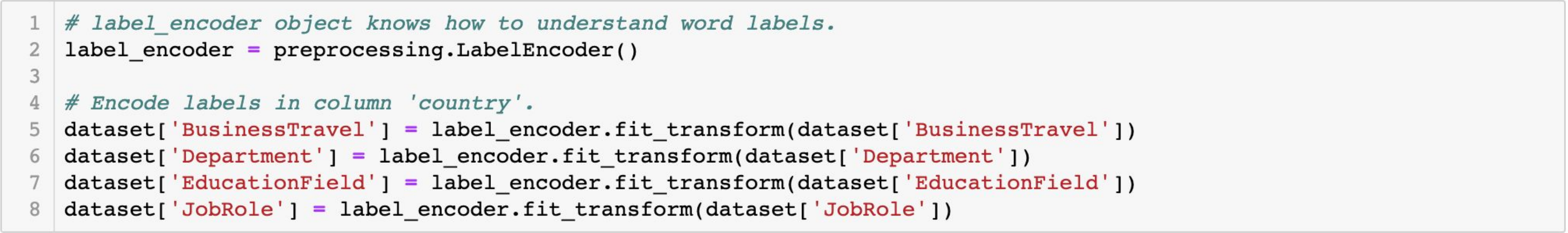

MonthlyIncome
20000
17500
15000
12500
10000
7500
5000
2500
0
Yes
No
OverTime
Attrition
Yes
No


MonthlyIncome
20000
17500
15000
12500
10000
7500
5000
2500
0
1
2
3
4
JobSatisfaction
Attrition
Yes
No


Yes
No
count
0
100
200
300
400
500
600
Single
Married
Divorced
MaritalStatus

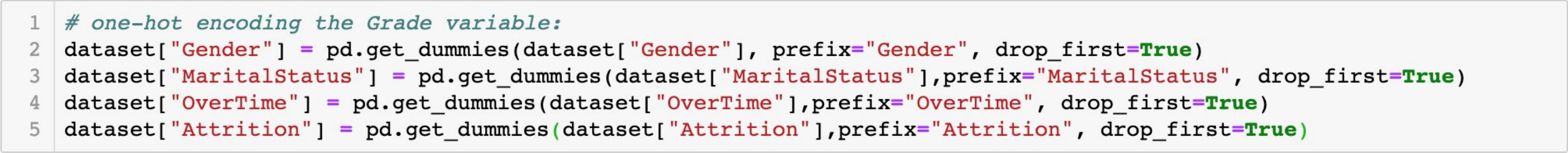

Yes
No
count
0
200
400
600
800
Travel_Rarely
Travel_Frequently
Non-Travel
BusinessTravel

Feature Encoding



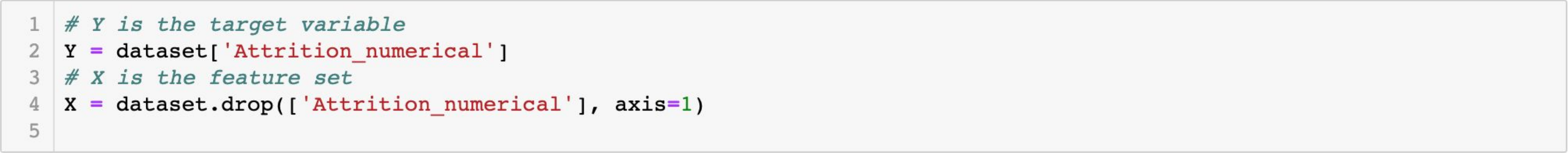
```python
# label_encoder object knows how to understand word labels.
label_encoder = preprocessing.LabelEncoder()

# Encode labels in column 'country'.
dataset['BusinessTravel'] = label_encoder.fit_transform(dataset['BusinessTravel'])
dataset['Department'] = label_encoder.fit_transform(dataset['Department'])
dataset['EducationField'] = label_encoder.fit_transform(dataset['EducationField'])
dataset['JobRole'] = label_encoder.fit_transform(dataset['JobRole'])
```



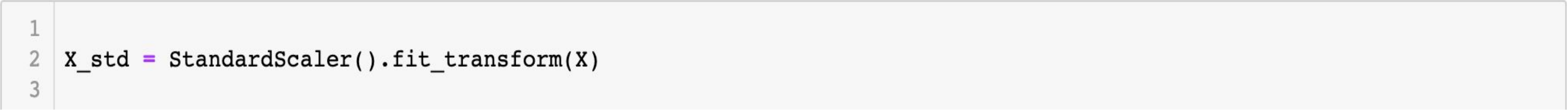
```python
# one-hot encoding the Grade variable:
dataset["Gender"] = pd.get_dummies(dataset["Gender"], prefix="Gender", drop_first=True)
dataset["MaritalStatus"] = pd.get_dummies(dataset["MaritalStatus"],prefix="MaritalStatus", drop_first=True)
dataset["OverTime"] = pd.get_dummies(dataset["OverTime"],prefix="OverTime", drop_first=True)
dataset["Attrition"] = pd.get_dummies(dataset["Attrition"],prefix="Attrition", drop_first=True)
```

Defining X and Y



```python
# Y is the target variable
Y = dataset['Attrition_numerical']
# X is the feature set
X = dataset.drop(['Attrition_numerical'], axis=1)

```

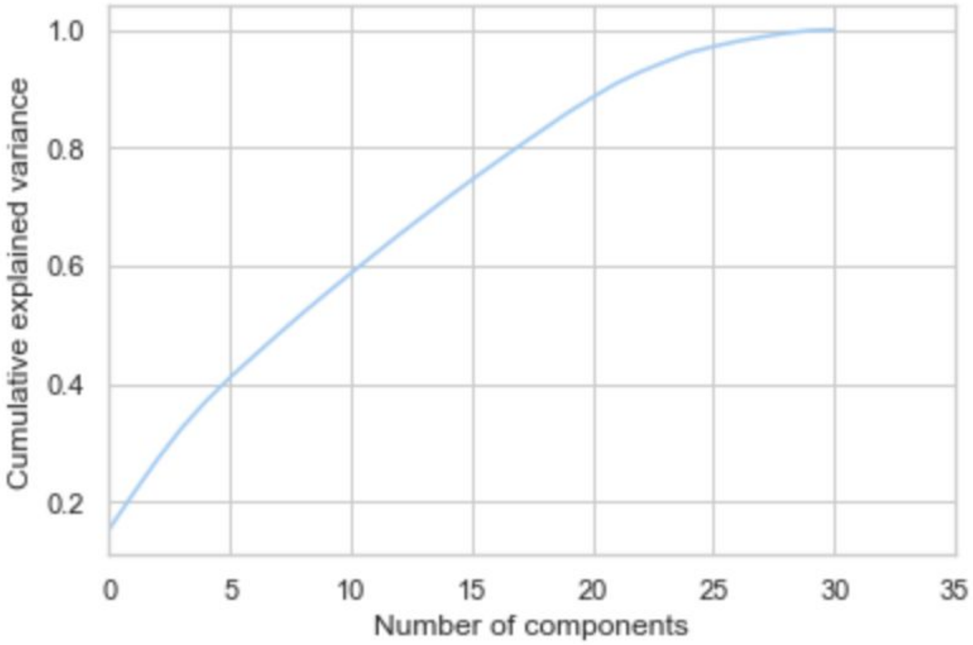


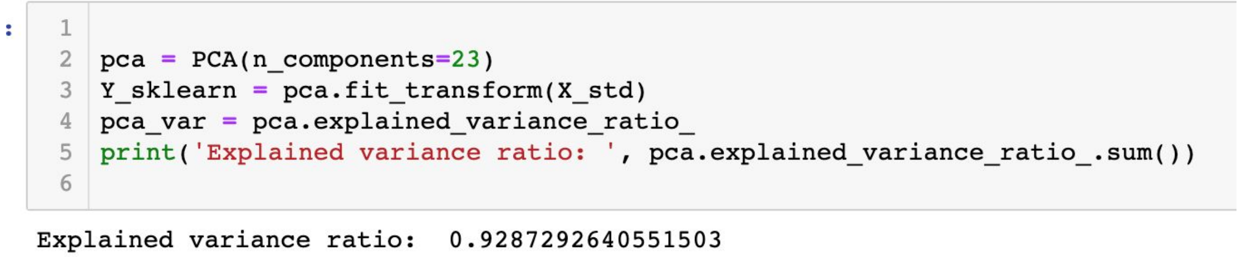
```python

X_std = StandardScaler().fit_transform(X)

```

PCA





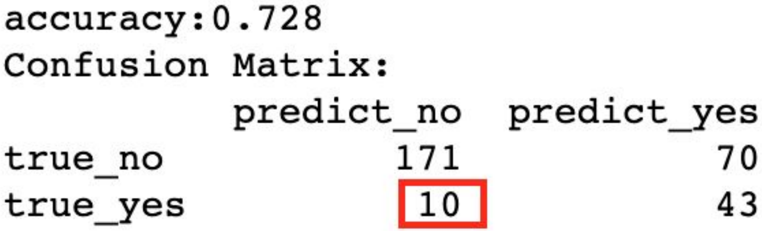
```

pca = PCA(n_components=23)
Y_sklearn = pca.fit_transform(X_std)
pca_var = pca.explained_variance_ratio_
print('Explained variance ratio: ', pca.explained_variance_ratio_.sum())

```

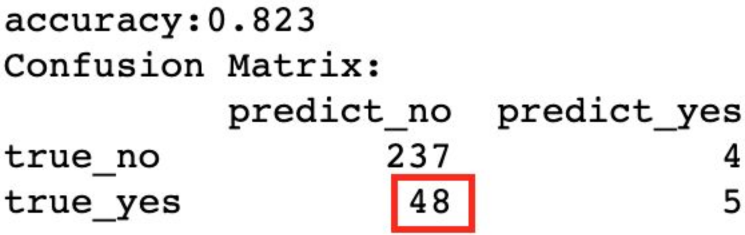
```
Explained variance ratio:  0.9287292640551503
```

Logistic Regression Classifier



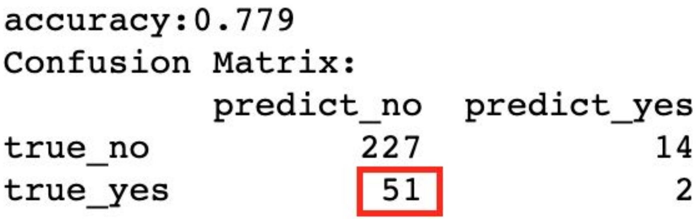
```
accuracy:0.728
Confusion Matrix:
         predict_no  predict_yes
true_no         171           70
true_yes         10           43
```

KNN



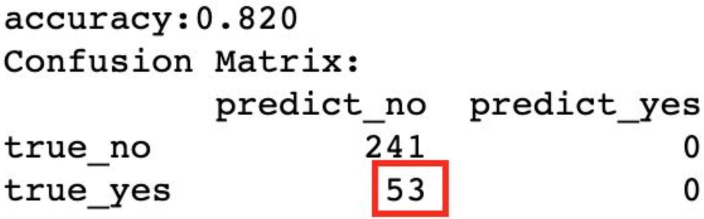
```
accuracy:0.823
Confusion Matrix:
         predict_no  predict_yes
true_no         237            4
true_yes         48            5
```

Decision Tree



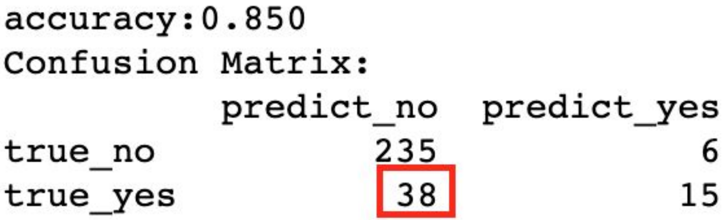
```
accuracy:0.779
Confusion Matrix:
         predict_no  predict_yes
true_no         227           14
true_yes         51            2
```

Random Forest



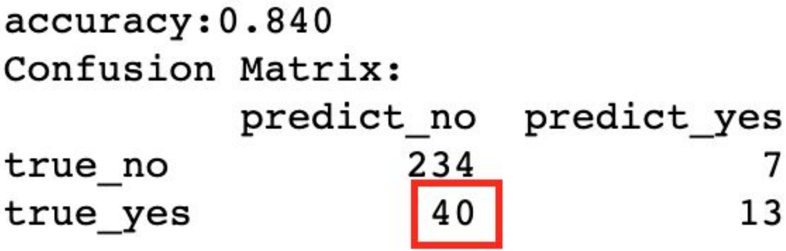
```
accuracy:0.820
Confusion Matrix:
         predict_no  predict_yes
true_no         241            0
true_yes         53            0
```

Gradient Boosting



```
accuracy:0.850
Confusion Matrix:
         predict_no  predict_yes
true_no         235            6
true_yes         38           15
```

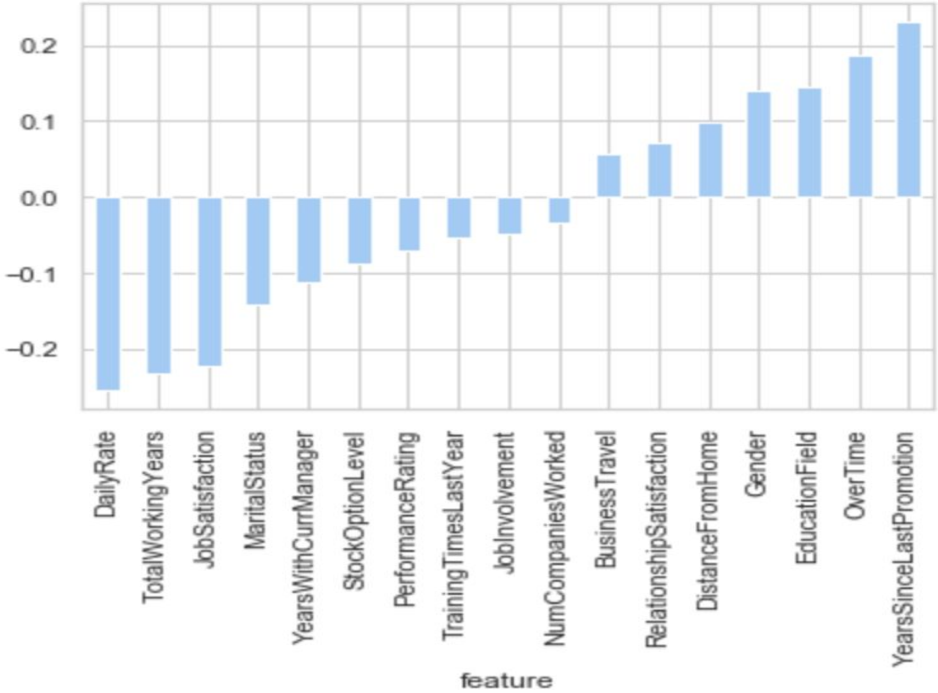
Naive Bayes



```
accuracy:0.840
Confusion Matrix:
         predict_no  predict_yes
true_no         234            7
true_yes         40           13
```

Feature Importance

- lr: is the fitted logistic regression and winning model
- list_of_feat: list of features (max coef) for each component



That's not direct feature! It’s max feature for this component.

CONCLUSION

Factors associated with higher attrition risks:

- working environment:
  - working overtime
  - living far away from the company
  - lack of satisfaction
  - raveling a lot
- job roles:
  - sales
  - HR
  - research scientists
  - laboratory technician

Also:

- junior level
- single
- young age

Future Ideas

- For employees who have been working overtime and traveling a lot - give compensation or time off

- Provide more team and culture building

- Additional attention (like mentorship) to junior employees

- Stress relieve like massage, gym or game room