Human Resources:
A Study of Attrition

...

Matt Gracer

Data

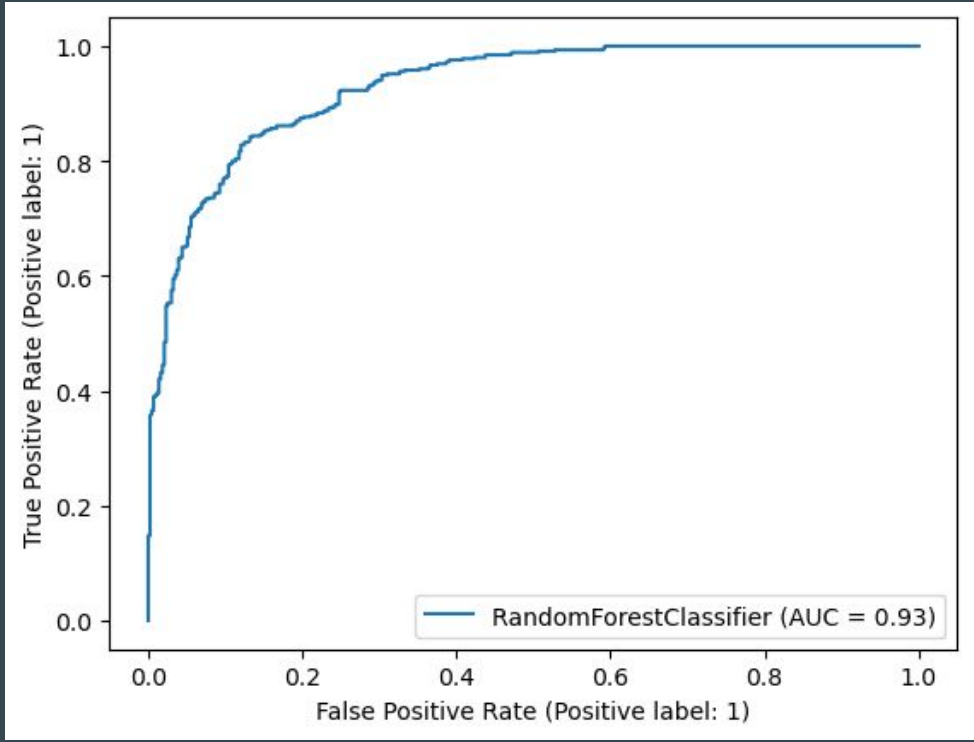
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1480 entries, 0 to 1479
Data columns (total 38 columns):
 #   Column                    Non-Null Count  Dtype 
---  ------                    --------------  ----- 
 0   EmpID                     1480 non-null   object
 1   Age                       1480 non-null   int64 
 2   AgeGroup                  1480 non-null   object
 3   Attrition                 1480 non-null   object
 4   BusinessTravel            1480 non-null   object
 5   DailyRate                 1480 non-null   int64 
 6   Department                1480 non-null   object
 7   DistanceFromHome          1480 non-null   int64 
 8   Education                 1480 non-null   int64 
 9   EducationField            1480 non-null   object
 10  EmployeeCount             1480 non-null   int64 
 11  EmployeeNumber            1480 non-null   int64 
 12  EnvironmentSatisfaction   1480 non-null   int64 
 13  Gender                    1480 non-null   object
 14  HourlyRate                1480 non-null   int64 
 15  JobInvolvement            1480 non-null   int64 
 16  JobLevel                  1480 non-null   int64 
 17  JobRole                   1480 non-null   object
 18  JobSatisfaction           1480 non-null   int64 
 19  MaritalStatus             1480 non-null   object
 20  MonthlyIncome             1480 non-null   int64 
 21  SalarySlab                1480 non-null   object
 22  MonthlyRate               1480 non-null   int64 
 23  NumCompaniesWorked        1480 non-null   int64 
 24  Over18                    1480 non-null   object
 25  OverTime                  1480 non-null   object
 26  PercentSalaryHike         1480 non-null   int64 
 27  PerformanceRating         1480 non-null   int64 
 28  RelationshipSatisfaction  1480 non-null   int64 
 29  StandardHours             1480 non-null   int64 
```

- ➢ 38 Columns
- ➢ 1480 Observations
- ➢ Gender, Marital Status, Monthly Income, and YearsinCurrentRole were all target variables

Data Organization

- Balance data
- One Hot Encode variables
- Train Test Set
- Models: Random Forest is the baseline model
- Additional Models: Logistic Regression, Decision Tree Classifier, and Multinomial
- Hyperparameters were added for more iterations and then higher tolerance with confusion matrix
- SMOTE function
- Linear Regression - Target variables were Monthly Income and YearsInCurrentRole
- Multi-class Regression - Target variable was Marital Status

Baseline Model - Random Forest Regressor



- Accuracy score is 0.93 on the train data

Additional Models - Classification Report

Model 1 – Accuracy: 0.5033783783783784
Classification Report:

	precision	recall	f1-score	support
0	0.55	0.45	0.49	159
1	0.47	0.57	0.51	137
accuracy			0.50	296
macro avg	0.51	0.51	0.50	296
weighted avg	0.51	0.50	0.50	296

Model 2 – Accuracy: 0.5236486486486487
Classification Report:

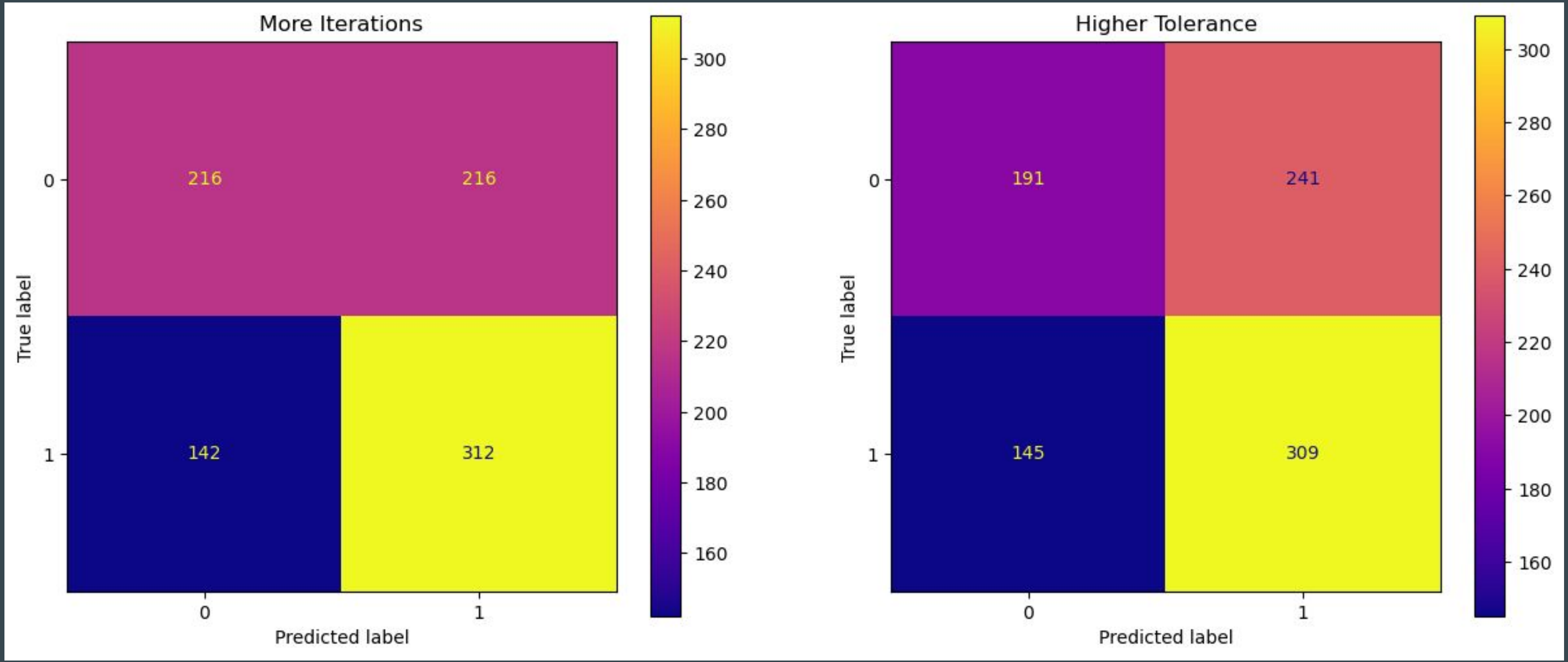
	precision	recall	f1-score	support
0	0.56	0.53	0.54	159
1	0.49	0.52	0.50	137
accuracy			0.52	296
macro avg	0.52	0.52	0.52	296
weighted avg	0.53	0.52	0.52	296

Model 3 – Accuracy: 0.543918918918919
Classification Report:

	precision	recall	f1-score	support
0	0.60	0.45	0.52	159
1	0.51	0.65	0.57	137
accuracy			0.54	296
macro avg	0.55	0.55	0.54	296
weighted avg	0.56	0.54	0.54	296

- Model 1 - Logistic Regression
- Model 2 - Decision Tree Classifier
- Model 3 - Multinomial
- Scores were lower for these models
- Baseline Random Forest Regressor Model had highest results

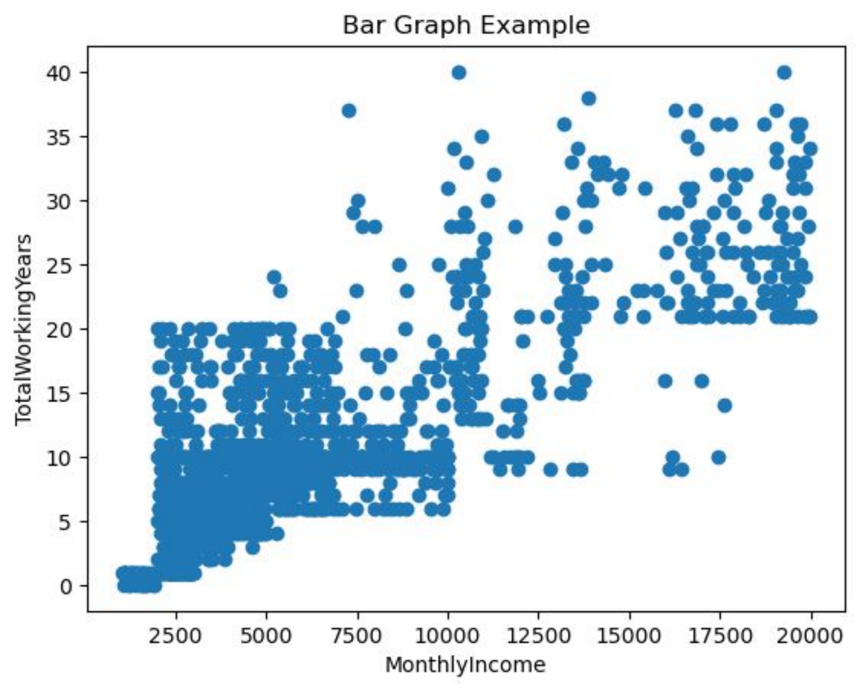
Hyperparameters for Logistic Regression



- ➢ Confusion Matrix for More Parameters and higher Tolerance

Scatter Plot for Linear Regression

- ➢ Monthly Income on X Axis
- ➢ TotalWorkingHours on Y Axis
- ➢ Positive Correlation



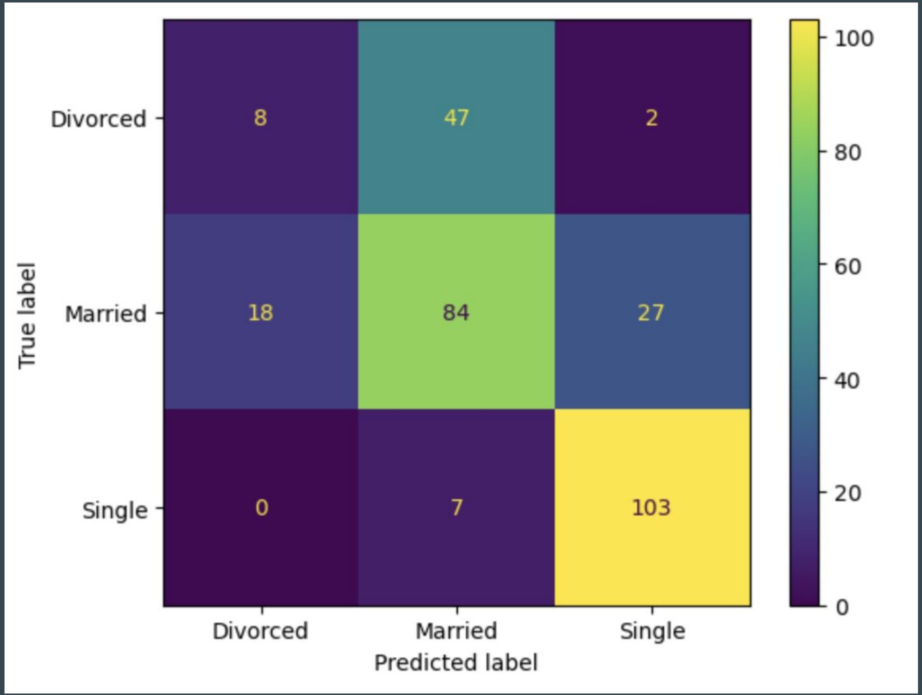
Linear Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          MonthlyIncome   R-squared:                       0.248
Model:                            OLS   Adj. R-squared:                  0.247
Method:                 Least Squares   F-statistic:                     388.2
Date:                Mon, 06 Nov 2023   Prob (F-statistic):           6.10e-75
Time:                        13:02:28   Log-Likelihood:                -11322.
No. Observations:                1182   AIC:                         2.265e+04
Df Residuals:                    1180   BIC:                         2.266e+04
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const           3306.6715    162.684     20.326      0.000    2987.490    3625.853
YearsAtCompany   384.5406     19.516     19.704      0.000     346.250     422.831
==============================================================================
Omnibus:                      371.198   Durbin-Watson:                   1.752
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1066.059
Skew:                           1.602   Prob(JB):                    3.22e-232
Kurtosis:                       6.373   Cond. No.                         13.4
==============================================================================
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          MonthlyIncome   R-squared:                       0.539
Model:                            OLS   Adj. R-squared:                  0.535
Method:                 Least Squares   F-statistic:                     171.1
Date:                Tue, 07 Nov 2023   Prob (F-statistic):          5.72e-191
Time:                        10:57:47   Log-Likelihood:                -11033.
No. Observations:                1182   AIC:                         2.208e+04
Df Residuals:                    1173   BIC:                         2.213e+04
Df Model:                           8
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                    1720.0221    803.853      2.140      0.033     142.872    3297.172
YearsAtCompany             50.2819     30.385      1.655      0.098      -9.333     109.897
YearsSinceLastPromotion    -5.7712     34.943     -0.165      0.869     -74.330      62.787
TrainingTimesLastYear     -18.2800     61.511     -0.297      0.766    -138.964     102.404
TotalWorkingYears         419.1931     15.541     26.974      0.000     388.703     449.684
PerformanceRating         -39.4628    221.465     -0.178      0.859    -473.974     395.048
MonthlyRate                 0.0154      0.011      1.368      0.172      -0.007       0.037
YearsInCurrentRole        -23.4194     38.020     -0.616      0.538     -98.014      51.175
HourlyRate                 -4.6900      3.956     -1.186      0.236     -12.451       3.071
==============================================================================
Omnibus:                       85.957   Durbin-Watson:                   2.050
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              166.524
Skew:                           0.485   Prob(JB):                     6.92e-37
Kurtosis:                       4.562   Cond. No.                     1.65e+05
==============================================================================
```

- Linear Regression for Monthly Income with one dependent variable
- Other Linear Regression is result with multiple dependent variables
- R**2 values varied - 0.25 and 0.539

Multi-Class Regression



	precision	recall	f1-score	support
Divorced	0.14	0.31	0.19	26
Married	0.65	0.61	0.63	138
Single	0.94	0.78	0.85	132
accuracy			0.66	296
macro avg	0.58	0.57	0.56	296
weighted avg	0.73	0.66	0.69	296

- Married and Single categories had highest scores

Recommendations

- Sort the data for specific genders to better understand how it is driving attrition rates - Random Forest Regressor had high scores on the train data
- The multi-Class Regression study had high scores for the specific categories of Single and Married so could filter the data for these categories to better understand how it is driving attrition rates
- For the Linear Regression Study filtering for the variables that were found to be significant at the 1% level would also be a solid indicator of what is driving attrition levels.
- These recommendations would be very helpful to companies because of how high levels of attrition are. Companies lose revenue when there’s high turnover so understanding how to interpret data for attrition is a big step toward raising revenue levels