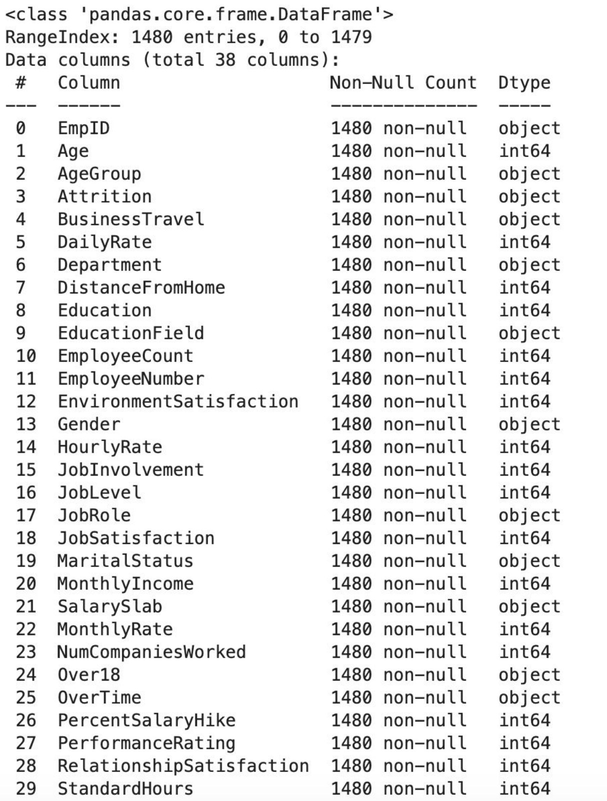
Human Resources: A Study of Attrition

•••

Matt Gracer

Data



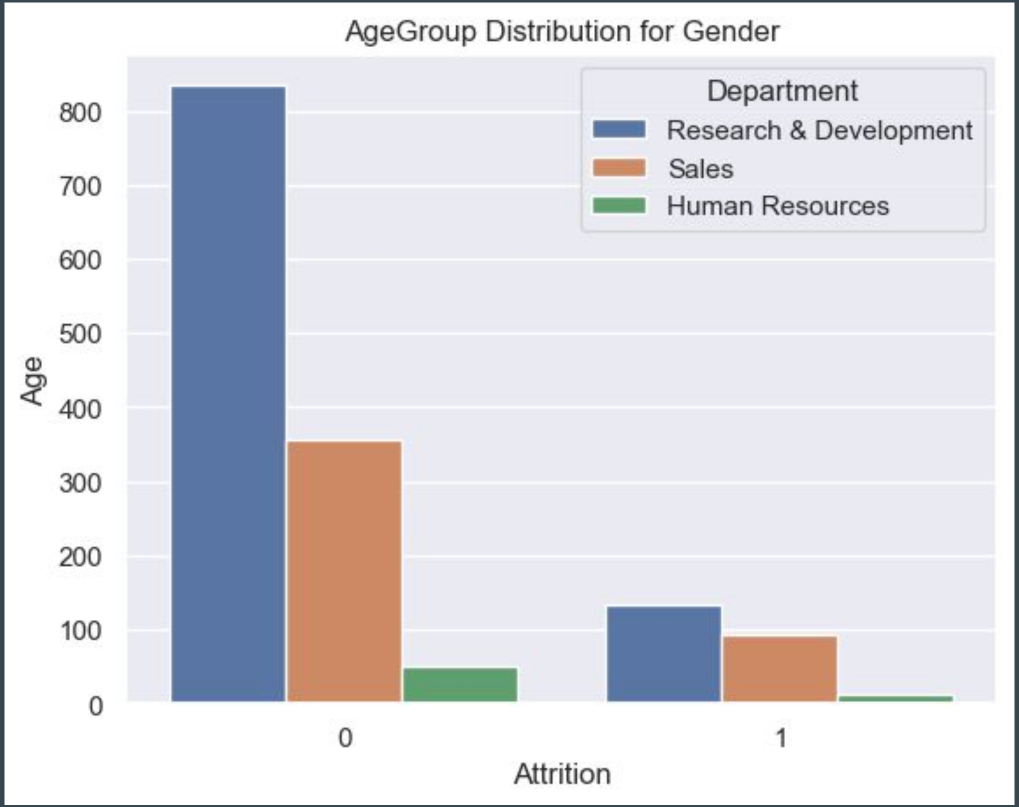
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1480 entries, 0 to 1479
Data columns (total 38 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   EmpID                     1480 non-null   object
 1   Age                       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
- Attrition, Gender, Marital Status, Monthly Income, and YearsinCurrentRole were all target variables

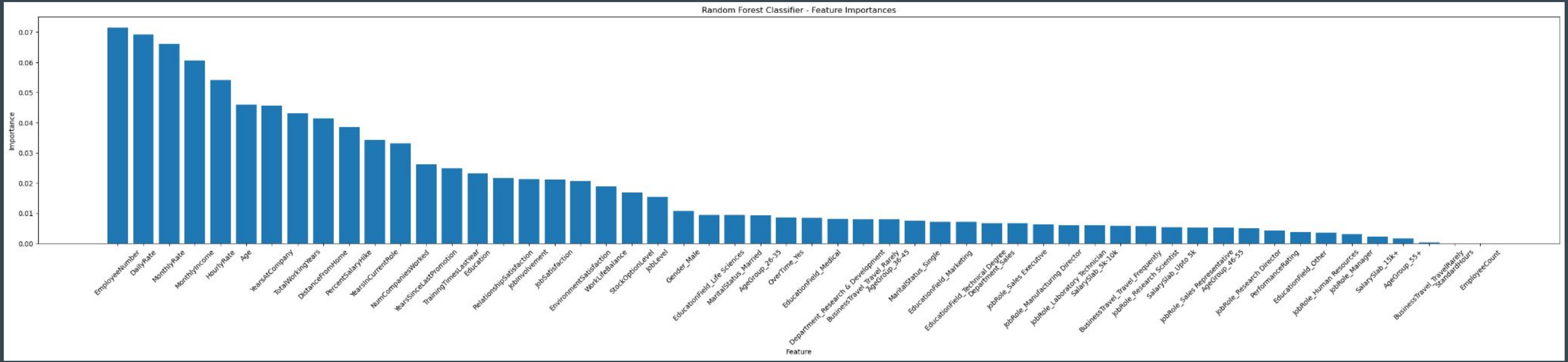
Data Organization

- Data was found from Kaggle
- Unbalanced Attrition had 1480 Observations, Balanced Attrition had 476
- 4 Total Models: Random Forest Classifier, Logistic Regression, Decision Tree Classifier, and Mulitnomial
- One Hot Encoder Function
- Merge datasets between numeric and Object
- Conducted for Train and Test Datasets
- Fit transformed train and test data on models
- Cross Validation for all models resulting in higher scores

Attrition Levels



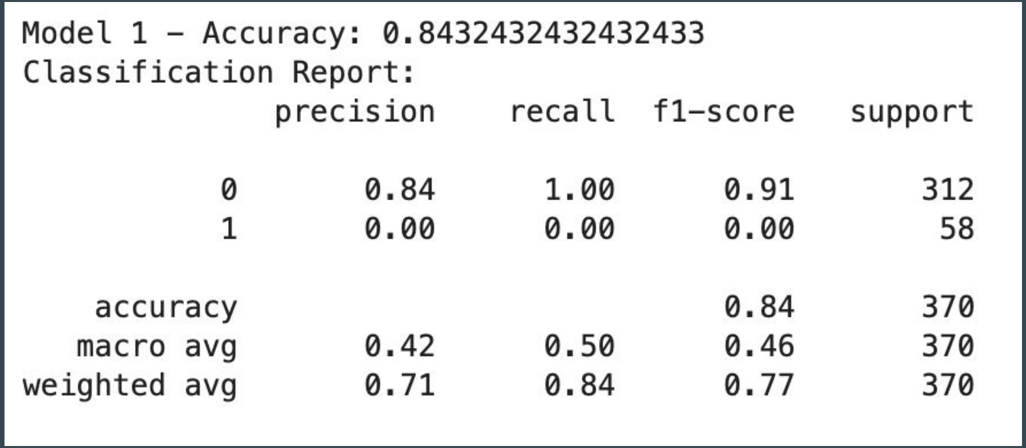
Feature Importances



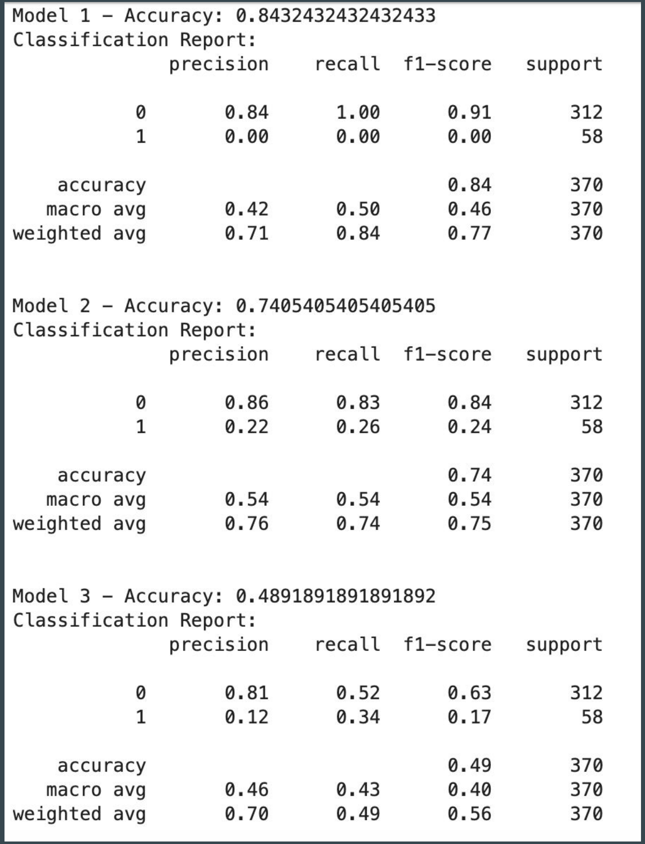
- ➢ Employee Number, Daily Rate, Monthly Rate, Monthly Income, Hourly Rate had highest Feature Importances

Additional Models - Classification Report Unbalanced

- Model 1 - Random Forest Classifier
- Model 2 - Logistic Regression
- Model 3 - Decision Tree Classifier
- Model 4 - Multinomial
- Scores were lower for these models
- Baseline Random Forest Regressor Model had highest results



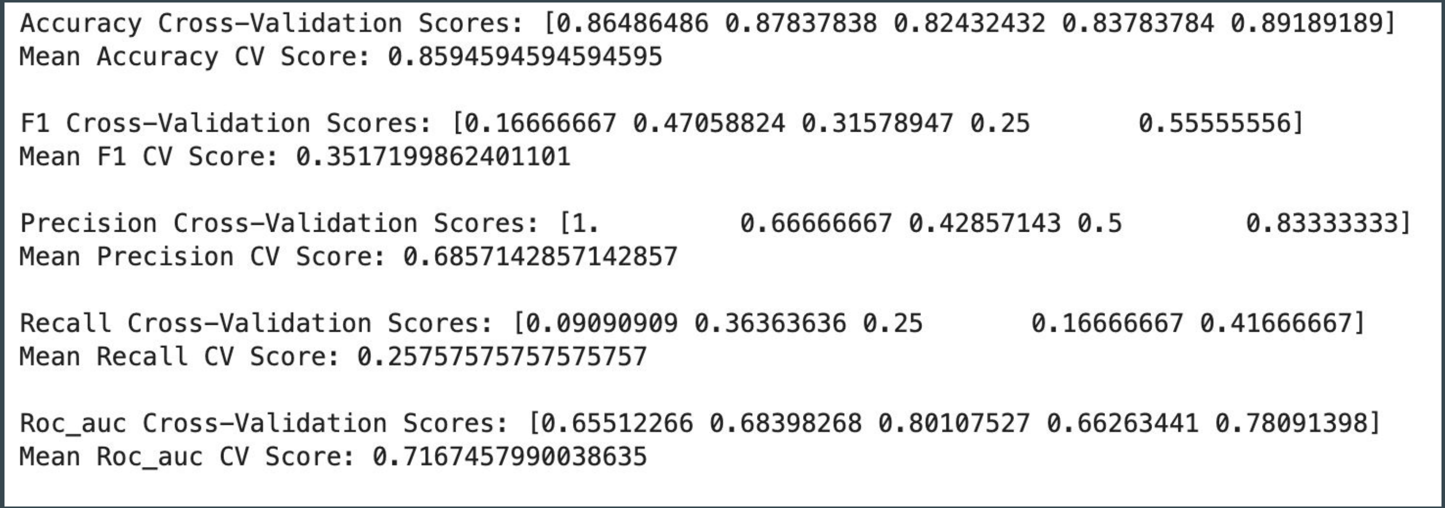
```
Model 1 - Accuracy: 0.8432432432432433
Classification Report:
              precision    recall  f1-score   support

           0       0.84      1.00      0.91       312
           1       0.00      0.00      0.00        58

    accuracy                           0.84       370
   macro avg       0.42      0.50      0.46       370
weighted avg       0.71      0.84      0.77       370
```



```
Model 1 - Accuracy: 0.8432432432432433
Classification Report:
              precision    recall  f1-score   support

           0       0.84      1.00      0.91       312
           1       0.00      0.00      0.00        58

    accuracy                           0.84       370
   macro avg       0.42      0.50      0.46       370
weighted avg       0.71      0.84      0.77       370
```

```
Model 2 - Accuracy: 0.7405405405405405
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.83      0.84       312
           1       0.22      0.26      0.24        58

    accuracy                           0.74       370
   macro avg       0.54      0.54      0.54       370
weighted avg       0.76      0.74      0.75       370


Model 3 - Accuracy: 0.4891891891891892
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.52      0.63       312
           1       0.12      0.34      0.17        58

    accuracy                           0.49       370
   macro avg       0.46      0.43      0.40       370
weighted avg       0.70      0.49      0.56       370
```

Cross Validation

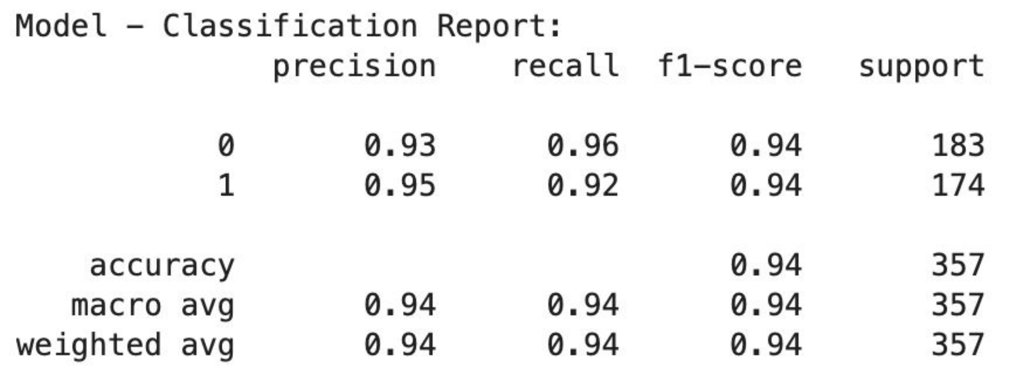


```
Accuracy Cross-Validation Scores: [0.86486486 0.87837838 0.82432432 0.83783784 0.89189189]
Mean Accuracy CV Score: 0.8594594594594595

F1 Cross-Validation Scores: [0.16666667 0.47058824 0.31578947 0.25       0.55555556]
Mean F1 CV Score: 0.3517199862401101

Precision Cross-Validation Scores: [1.         0.66666667 0.42857143 0.5        0.83333333]
Mean Precision CV Score: 0.6857142857142857

Recall Cross-Validation Scores: [0.09090909 0.36363636 0.25       0.16666667 0.41666667]
Mean Recall CV Score: 0.25757575757575757

Roc_auc Cross-Validation Scores: [0.65512266 0.68398268 0.80107527 0.66263441 0.78091398]
Mean Roc_auc CV Score: 0.7167457990038635
```

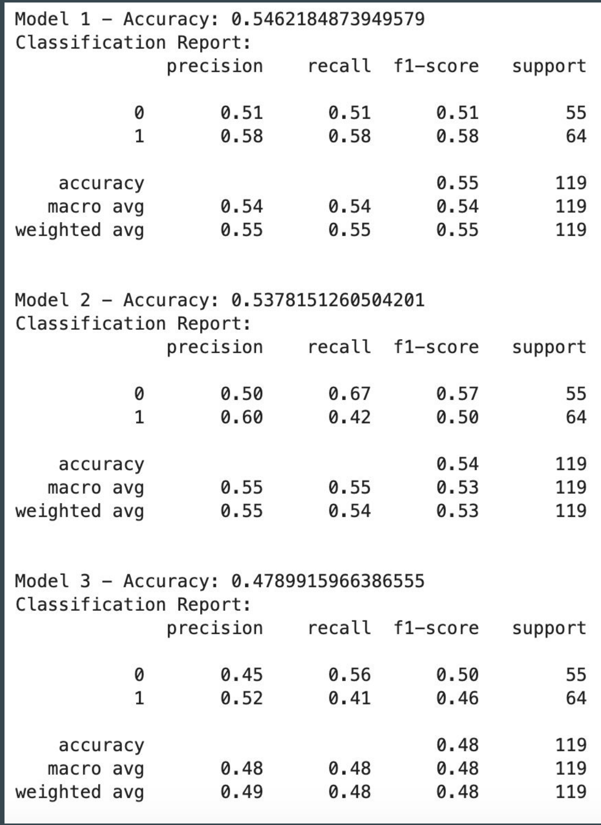
- ➢ Mean Accuracy remains around 85%
- ➢ Mean Precision increases to 68% and mean ROC increases to 71% from 53%

Classification Models - Balanced

- Model 1 - Random Forest Classifier
- Model 2 - Logistic Regression
- Model 3 - Decision Tree Classifier
- Model 4 - Multinomial
- Scores were lower for these models
- Baseline Random Forest Regressor Model had highest results

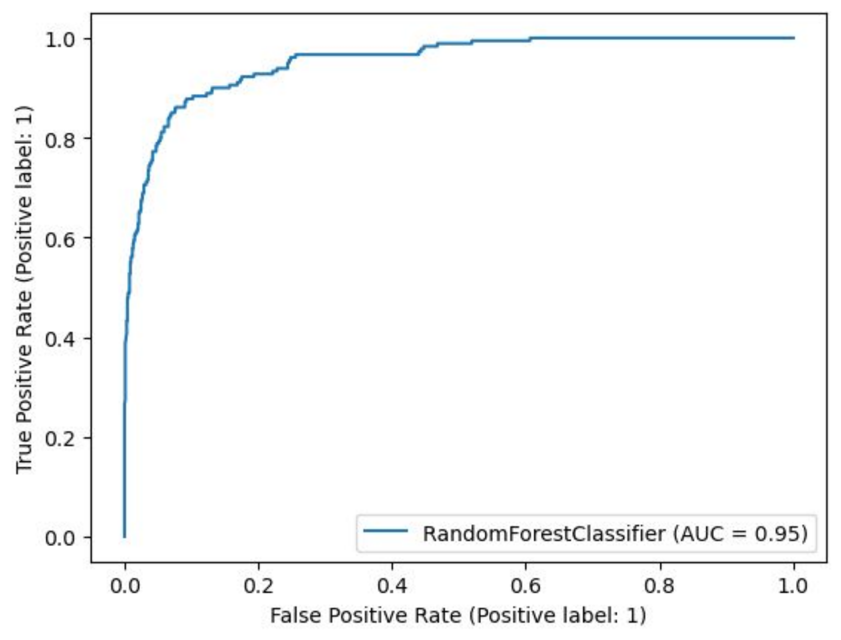


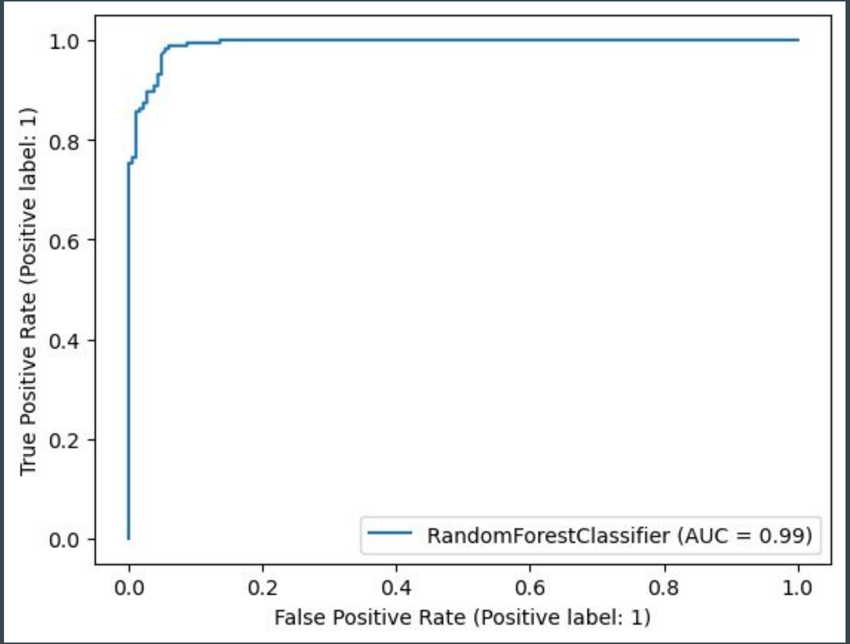
```
Model – Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.96      0.94       183
           1       0.95      0.92      0.94       174

    accuracy                           0.94       357
   macro avg       0.94      0.94      0.94       357
weighted avg       0.94      0.94      0.94       357
```



```
Model 1 – Accuracy: 0.5462184873949579
Classification Report:
              precision    recall  f1-score   support

           0       0.51      0.51      0.51        55
           1       0.58      0.58      0.58        64

    accuracy                           0.55       119
   macro avg       0.54      0.54      0.54       119
weighted avg       0.55      0.55      0.55       119


Model 2 – Accuracy: 0.5378151260504201
Classification Report:
              precision    recall  f1-score   support

           0       0.50      0.67      0.57        55
           1       0.60      0.42      0.50        64

    accuracy                           0.54       119
   macro avg       0.55      0.55      0.53       119
weighted avg       0.55      0.54      0.53       119


Model 3 – Accuracy: 0.4789915966386555
Classification Report:
              precision    recall  f1-score   support

           0       0.45      0.56      0.50        55
           1       0.52      0.41      0.46        64

    accuracy                           0.48       119
   macro avg       0.48      0.48      0.48       119
weighted avg       0.49      0.48      0.48       119
```

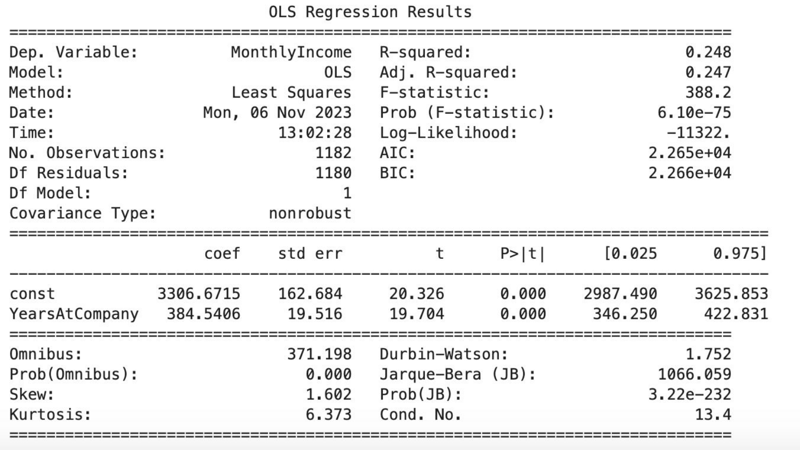
Baseline Model - Random Forest Classifier



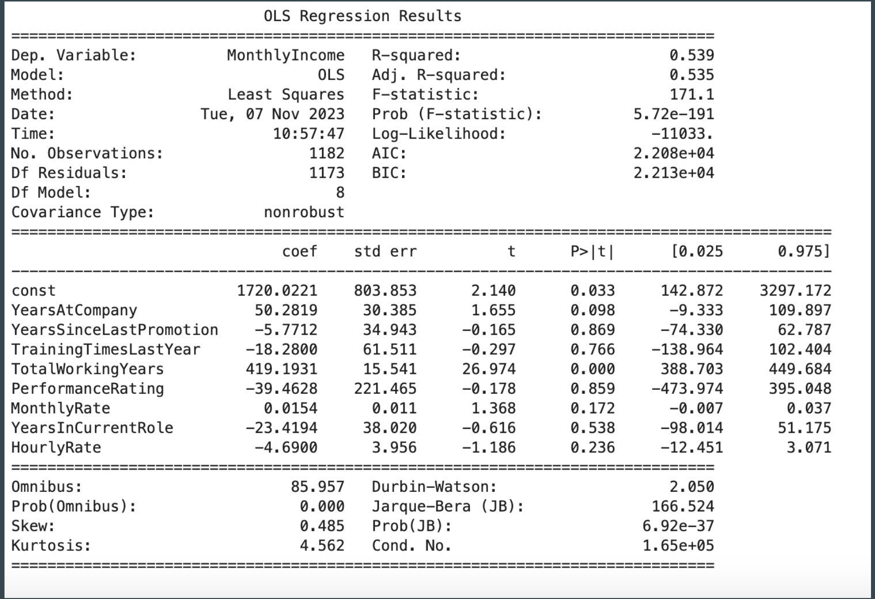


- Roc Curve for Unbalanced dataset was 0.95
- Roc Curve for Balanced dataset was 0.99

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

Recommendations

- The strongest model for Attrition was the Random Forest Classifier. The metrics were most reliable for the balanced Attrition study so the initial recommendation is to focus on that balanced dataset for Attrition to further understand how to keep Attrition low for those employees.
- 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 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

Thank You!