

Employee Attrition Data Mining **IBM HR Dataset**

Data Mining (MSCA 31008) Final Project

Nic Carlson, Martin Copello, Michael Jason, Matt Kendall, Nick Lesh



Agenda

- Problem Statement & Project Overview
- Data Assumptions and Challenges
- Clustering
- Linear Classifiers
- Non-Linear Classifiers
- Model Comparison
- Business Conclusions

Problem Statement & Project Overview

Gallup states that attrition costs US companies at least one trillion dollars annually!

Our objectives with this project:

- Study data for employees who both leave and stay to gain a better understanding of factors that contribute to attrition
- Attempt to develop a reasonable strategy for companies to minimize attrition

We studied the publicly available IBM HR dataset, available via Kaggle.

Some questions we looked to answer included:

- Which company-controlled factors determine who is most likely to leave the company (also known as churners)?
- Does job satisfaction determine attrition or job performance?
- Which features make employees the happiest?
- Can we accurately predict which employees are likely to leave?

Data Assumptions and Challenges

Assumptions

General

Dataset reflects an entire organization (not filtered to specific departments, job levels, etc.)

Variable Interpretations

Attrition only represents voluntary departures from the organization

OverTime is a flag for employees that are eligible to take overtime (e.g. non-salaried employees) rather than employees that have actually logged overtime hours.

Dropped Variables

HourlyRate, DailyRate, MonthlyRate - Assuming that these are compensation related variables, although the interpretation is not clear in the documentation. Opting to use MonthlyIncome instead.

StandardHours, Over18 - No variation within these features

EmployeeCount, EmployeeNumber - Unique identifiers for each record

Challenges

Dimensionality

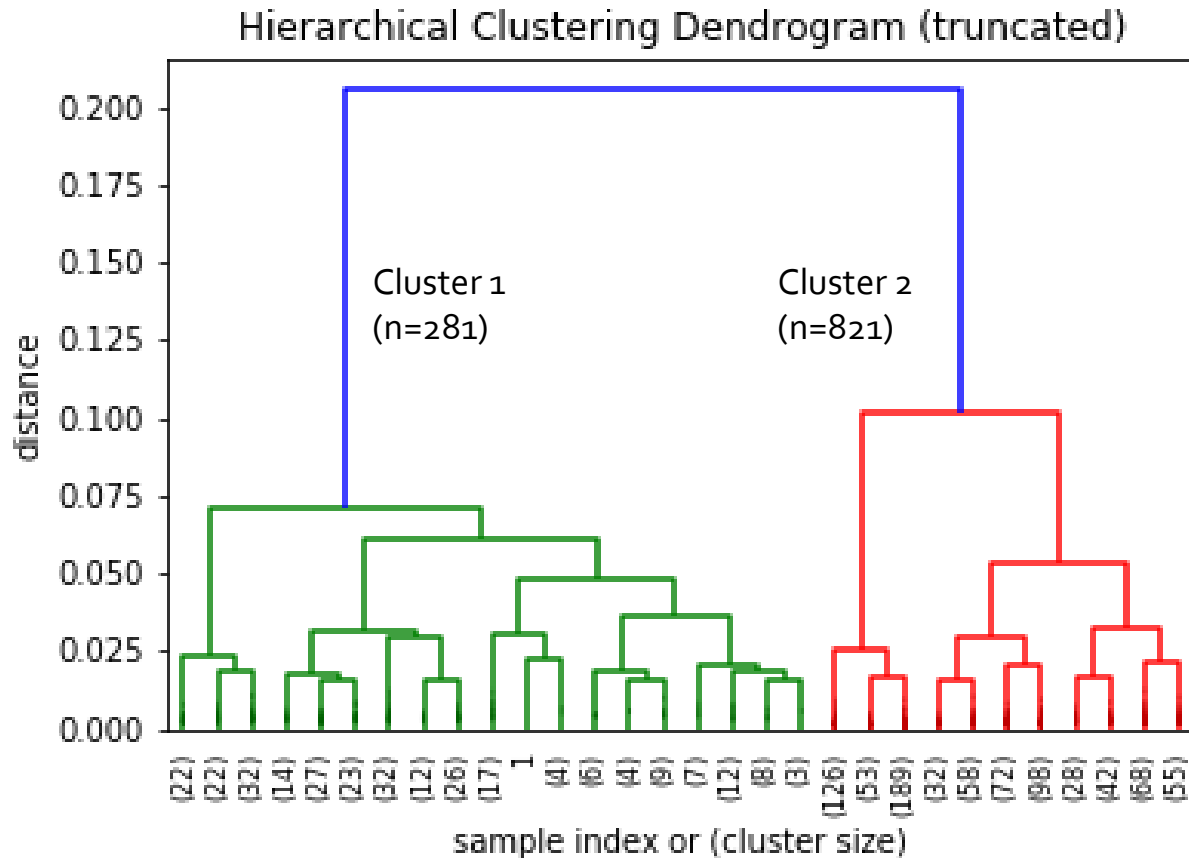
Little benefit from employing dimensionality reduction techniques (29 of ~45 features needed to capture 90% of the variance)

Imbalanced Target Variable (1:5)

The positive level in our target variable is only ~15% of all records, in response we:

- 1) Focused on binary **F1 Score** for target class instead of accuracy.
 - a) AUC, Confusion Matrix also considered
- 2) Applied **Synthetic Minority Oversampling Technique (SMOTE)** to our data.
- 3) For linear models, moved threshold for prediction to decrease Type I errors

Clustering and Dimensionality Reduction



Cluster 1 are generally characterized by...

- Younger employees, early in their careers
- That have less tenure and lower job levels in the company
- That received less training in the previous year and haven't recently been promoted (which makes sense if they weren't there in the previous year)
- They also live slightly further from the office

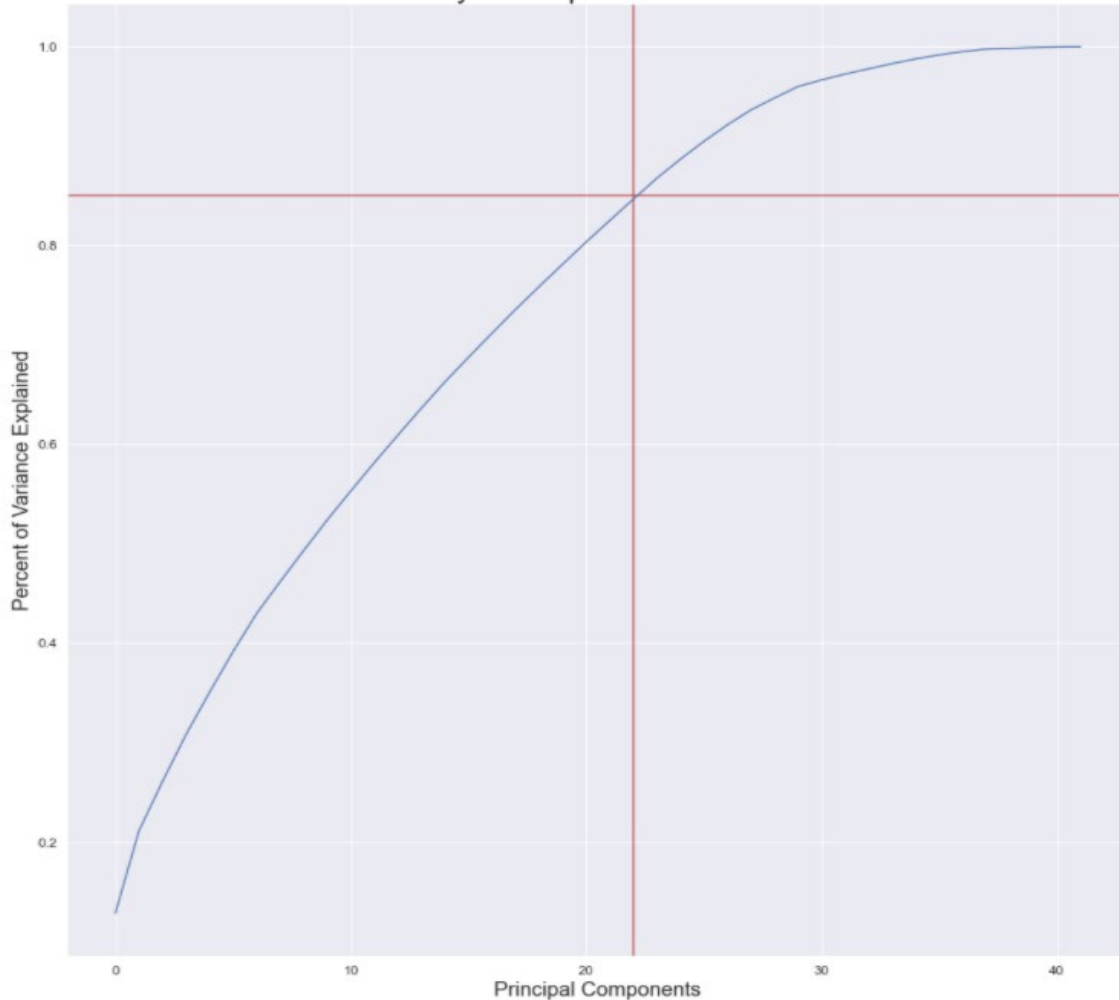
This was helpful to get a sense of what patterns to look for in our models

This paints a pretty vivid picture of cluster 1, and the similarity in features also explains difference in distance



Clustering and Dimensionality Reduction

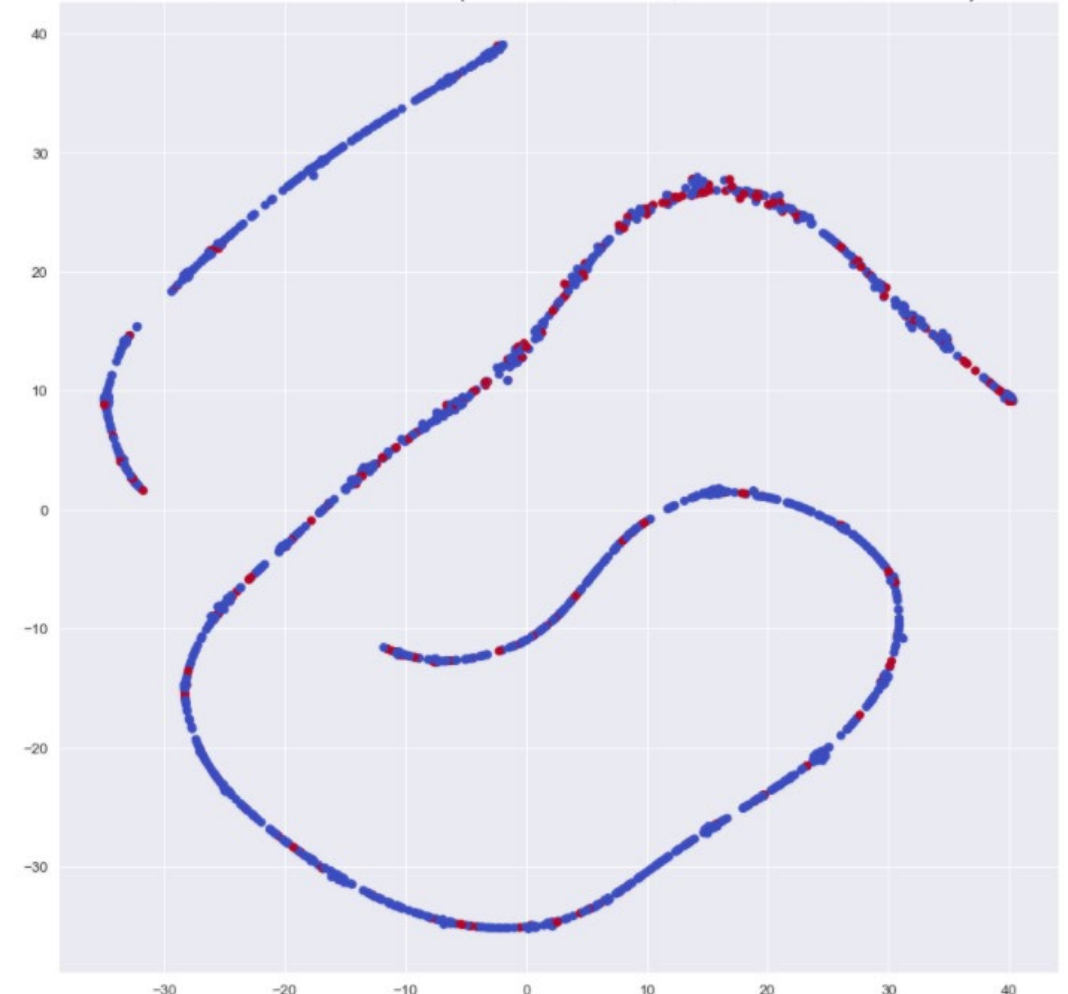
PCA Analysis - Explained Variance Ratio



2 components explain only 21.09% of variance. Need 24 components to explain over 85% of the variance. No large benefit to dimension reduction using PCA.

VS

t-SNE on IBM Dataset (Red = Churner, Blue = Non-Churner)



T-SNE breaks data into 2 to 3 clusters/lines, however churners seem clustered with non-churners. No discernable cluster for churners vs non-churners

Linear Classifiers

3 linear classification models created

Support vector classifier

Unregularized logistic regression
(statsmodel library)

Regularized logistic regression
(sklearn library)

Validation

F1 Score, ROC Curves, AUC,
Confusion Matrix

Data Imbalance

Threshold selection

Variable Selections

Dropped over half the columns for
parsimony and avoiding overfitting
(20 df in final regression models)

Added a quadratic age term to
capture impact of retirement

Logit Regression Results

Dep. Variable: Attrition

No. Observations: 1102

Model: Logit

Df Residuals: 1081

Method: MLE

Df Model: 20

Date: Tue, 25 Aug 2020

Pseudo R-squ.: 0.2575

Time: 21:26:06

Log-Likelihood: -361.82

converged: True

LL-Null: -487.29

Covariance Type: nonrobust

LLR p-value: 7.393e-42

coef

std err

z

P>|z|

[0.025

0.975]

Age

-0.0346

0.044

-0.787

0.431

-0.121

0.052

DistanceFromHome

0.0422

0.012

3.648

0.000

0.020

0.065

EnvironmentSatisfaction

-0.2627

0.087

-3.017

0.003

-0.433

-0.092

Gender

0.3149

0.198

1.588

0.112

-0.074

0.704

JobInvolvement

-0.3283

0.128

-2.563

0.010

-0.579

-0.077

JobSatisfaction

-0.2957

0.087

-3.404

0.001

-0.466

-0.125

NumCompaniesWorked

0.1276

0.041

3.100

0.002

0.047

0.208

OverTime

1.8272

0.202

9.053

0.000

1.432

2.223

RelationshipSatisfaction

-0.1685

0.088

-1.906

0.057

-0.342

0.005

TrainingTimesLastYear

-0.1126

0.077

-1.467

0.142

-0.263

0.038

WorkLifeBalance

-0.2122

0.129

-1.643

0.100

-0.465

0.041

YearsAtCompany

0.0456

0.034

1.325

0.185

-0.022

0.113

YearsInCurrentRole

-0.1555

0.050

-3.085

0.002

-0.254

-0.057

YearsSinceLastPromotion

0.1338

0.044

3.056

0.002

0.048

0.220

YearsWithCurrManager

-0.1071

0.050

-2.143

0.032

-0.205

-0.009

BusinessTravel_Travel_Frequently

2.2197

0.491

4.525

0.000

1.258

3.181

BusinessTravel_Travel_Rarely

1.5141

0.458

3.307

0.001

0.617

2.411

Department_Research & Development

0.0253

0.501

0.050

0.960

-0.957

1.007

Department_Sales

0.6615

0.510

1.298

0.194

-0.338

1.661

MaritalStatus_Single

1.1059

0.199

5.545

0.000

0.715

1.497

age_sq

-0.0002

0.001

-0.297

0.766

-0.001

0.001

0

278

31

1

24

35

0

1

Non-Linear Classifiers (3 applied + methods to balance data)

Decision Tree

Performed most effectively out of the three

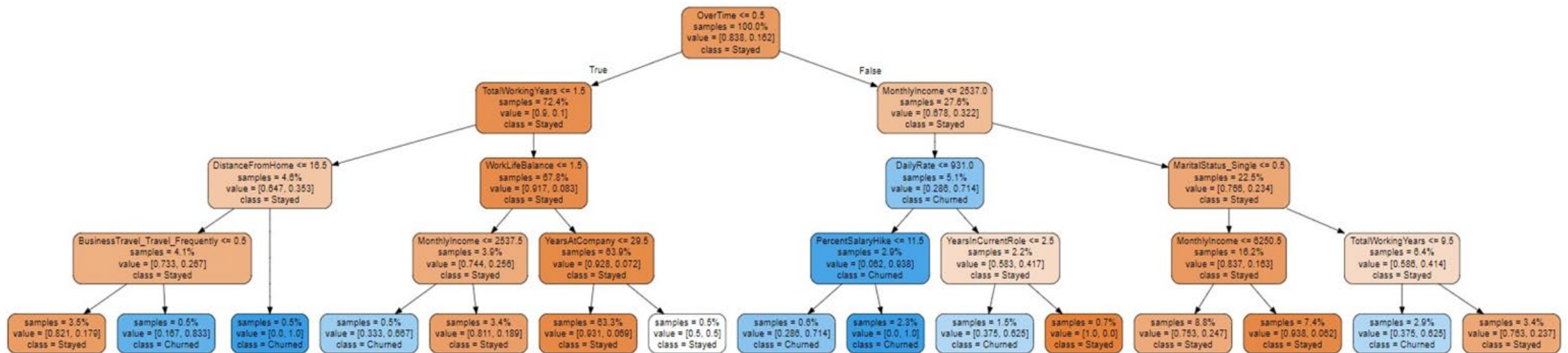
Lowered min_samples per leaf to allow for greater specificity given data imbalance

Bagging

Applied oversampling via BalancedBaggingClassifier to address imbalance, though it did not improve performance substantially

Random Forest

Similarly, applied balanced and bootstrap class weighting, but with unfavorable results

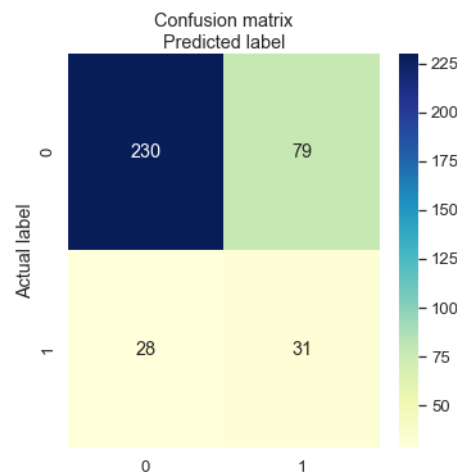


Non-Linear Classifiers (3 applied + methods to balance data)

Decision Tree

Performed most effectively out of the three

Lowered min_samples per leaf to allow for greater specificity given data imbalance



Bagging

Applied oversampling via `BalancedBaggingClassifier` to address imbalance, though it did not improve performance substantially

	precision	recall	f1-score	support
0	0.87	0.99	0.93	309
1	0.86	0.20	0.33	59
accuracy			0.87	368
macro avg	0.86	0.60	0.63	368
weighted avg	0.87	0.87	0.83	368

Random Forest

Similarly, applied balanced and bootstrap class weighting, but with unfavorable results



Non-Linear Classifiers (3 applied + methods to balance data)

Decision Tree

Performed most effectively out of the three

Lowered min_samples per leaf to allow for greater specificity given data imbalance






Bagging

Applied oversampling via `BalancedBaggingClassifier` to address imbalance, though it did not improve performance substantially

Random Forest

Similarly, applied balanced and bootstrap class weighting, but with unfavorable results

Model Comparison

Model	Logistic Regression	Support Vector Machine	Decision Tree	Bagged Decision Tree (Oversampling)	Random Forest (balanced sample class_weight)
F1 Score (binary) W/ SMOTE	0.56 0.45	0.51 0.40	0.34 0.37	0.33 (0.43) 0.43	0.32 (0.29) 0.31
Rating					
Top 3 attrition factors w/ best method	1. OverTime 2. Age 3. Marital Status	1. Overtime 2. MaritalStatus_Single 3. Business Travel - Travel Frequently	1. MonthlyIncome 2. OverTime 3. TotalWorkingYears	1. MonthlyIncome 2. DistnaceFromHome 3. Age	1. MonthlyIncome 2. Age 3. OverTime
Comments	Big win for parsimonious models in our overview. Notable here is that MonthlyIncome was excluded from the feature set, while it played a key role in the non-linear models.		Most effective of basic (non-hyper parameterized) Non-Linear models	Balanced bagging classifier did best for non-linear models overall, tied with using SMOTE data (performed poorly together)	Performed more poorly. Potentially overfit to training data

Conclusions

Apply data mining methods on fictitious IBM employee data to derive answers to the following questions:

- Understand which employees are leaving, in terms of characteristics and experiences
 - Per Logistic Regression, our best performing model, the characteristics that had the greatest predictor of an individual leaving included:
 - Eligibility for Overtime (i.e. Salary vs Hourly Employee)
 - Age
 - Marital Status
 - The data indicated younger, non-married, hourly employees tend to leave at greater rates than other individuals.
- Predict which current employees are at risk of leaving
 - We could apply these models to datasets of current employees to better predict what departments have the highest risk of employees leaving. While it would be difficult to hire only older, married, salaried individuals for future roles, there could be incentives such as stock options and other performance goals to entice high performing individuals to stay longer.
 - Correlation vs causality: Employees that were “Grandfathered in” in terms of benefits
- How we might specifically retain top performers (i.e. Is there anything unique about them? Can we capture that extra cost in our model?)
 - Current data states that 85% of employees received 3 out of 4 (“Excellent”) while 15% received 4 out of 4 (“Outstanding”).
 - Since there wasn’t great variability in performance rating, it is difficult to determine unique characteristics of top performers.
 - Either company is great at finding high performing individuals, or managers will have to assign scores differently (e.g. ranking system) to better identify true high performing individuals.
- Determine which levers the company can use to reduce attrition going forward (particularly for top performers)
 - As we were unable to identify true top performers, we will have to rely on the characteristics discovered from the logistic regression model to best reduce attrition going forward for all employees.

Possible Extensions

1. Additional Data Sources

- Real data
- Engagement survey
 - Anonymity as a constraint
- Granularity of performance management assessments
- Employee interaction data
- Benchmarking against other companies
- Feedback comments in official company portals
- Must be careful. Awareness of employment law or using health data

1. Additional analysis

- NLP and linguistic applications
 - Ethics and privacy concerns
- Graph Analyses
- Recommendation Algorithms

1. Implementation and Considerations

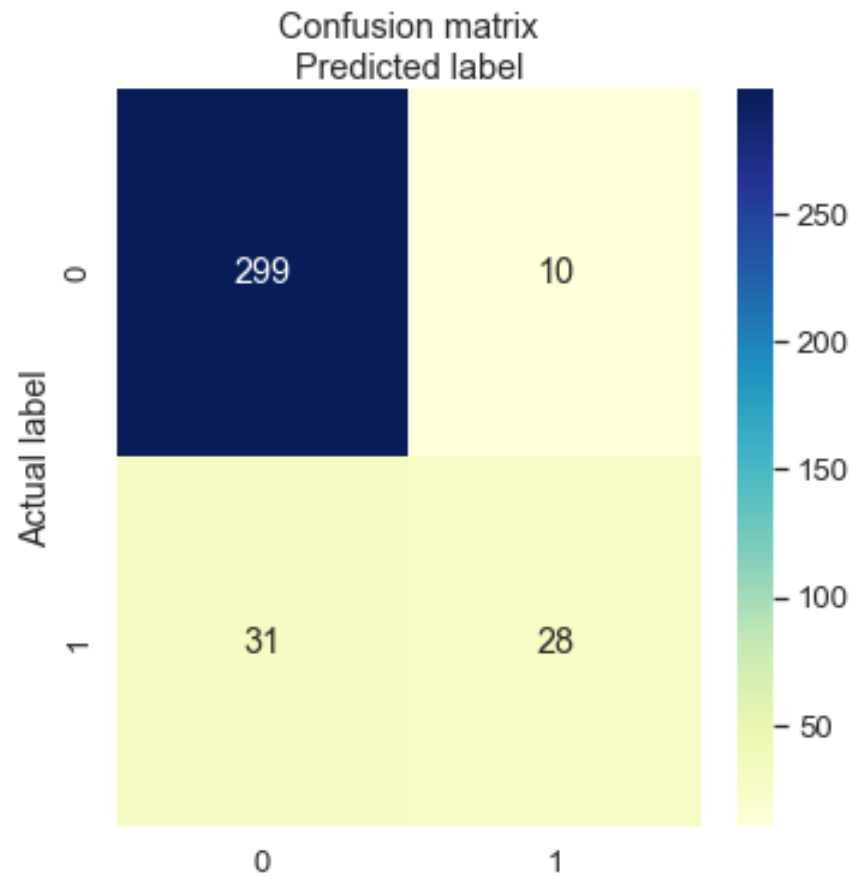
- Cost-benefit analysis of certain interviews/data collection methods
- Implementing new and innovative recruitment strategies
- Career ladder structure

Questions?

Appendix

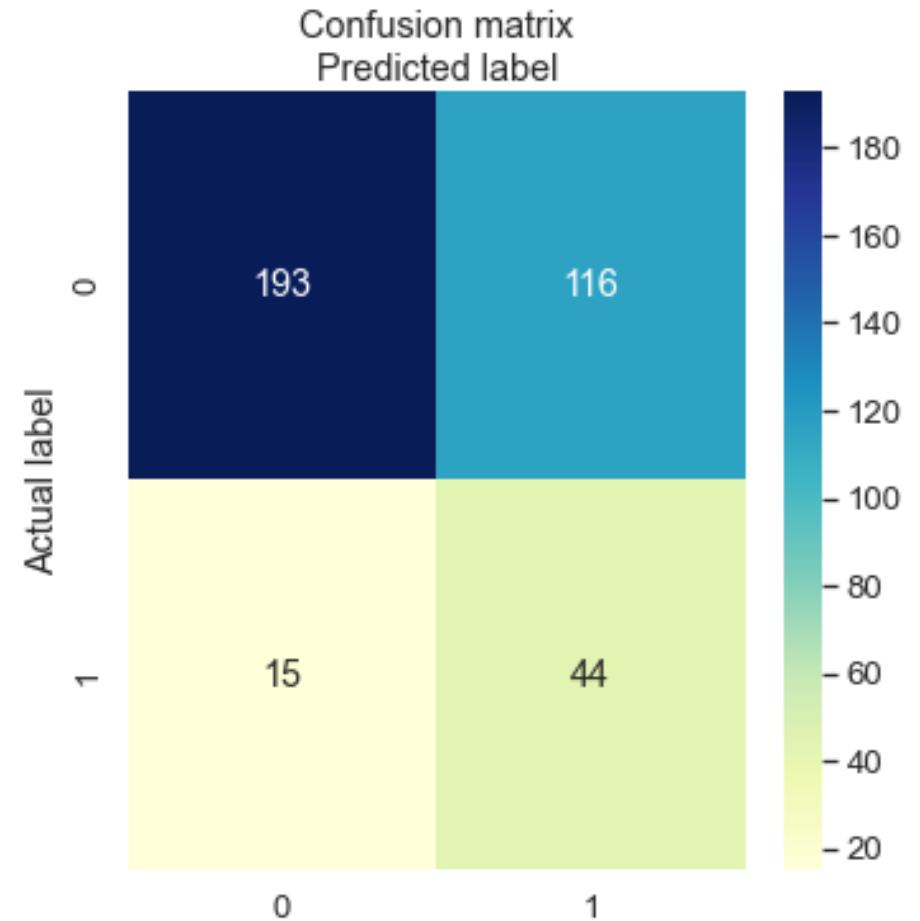
Confusion Matrix Considerations

Logistic Regression with SMOTE Data



F1 Score: 0.58

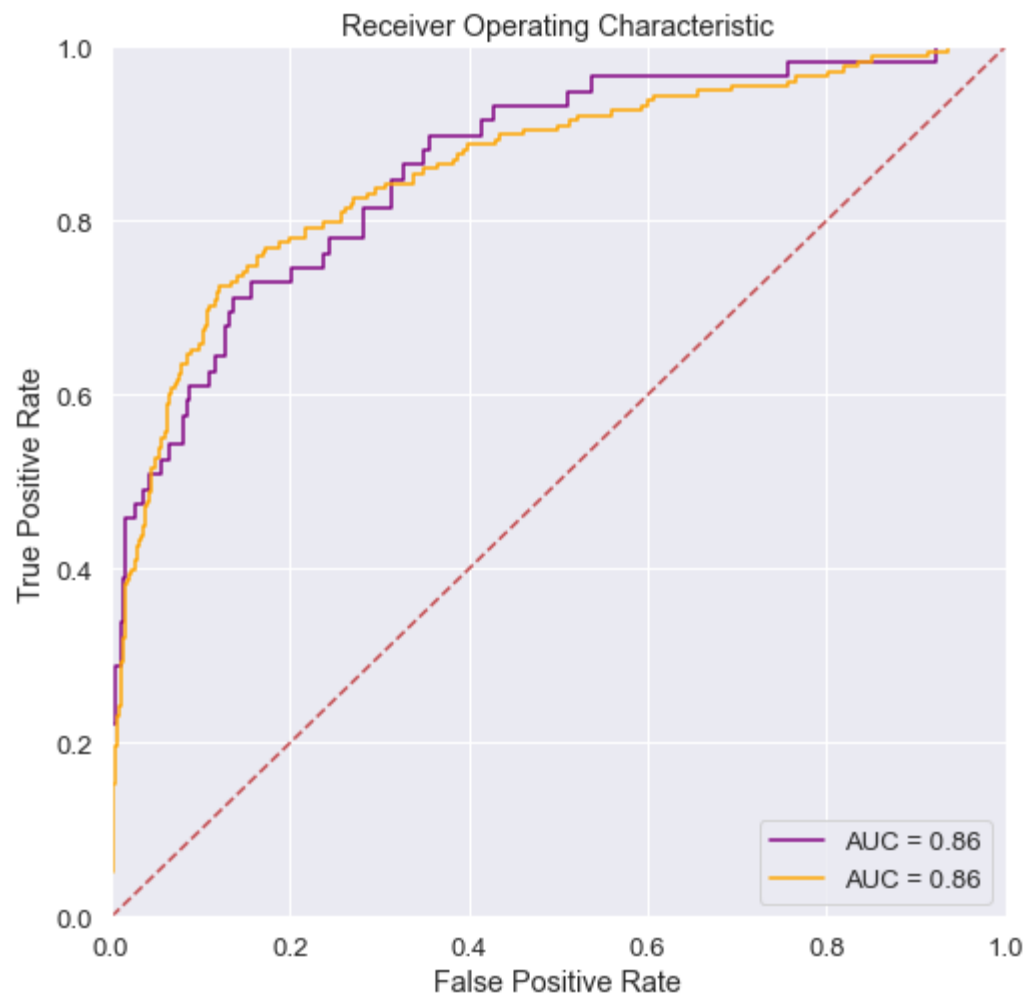
SVM with SMOTE Data



F1 Score: 0.40

Evidence of Overfitting

Logistic Regression



Logistic Regression with SMOTE Data

