# Employee Attrition predection

By: Muneera Alshunaifi

# Outline

Introduction

Data

Pre-processing

Modeling and evaluation

Conclusions

# Introduction

## WHAT IS ATTRITION

The Employee attrition occurs when an employee departs a company for a variety of reasons

## WHY?

Many reasons such as monthly income or job role

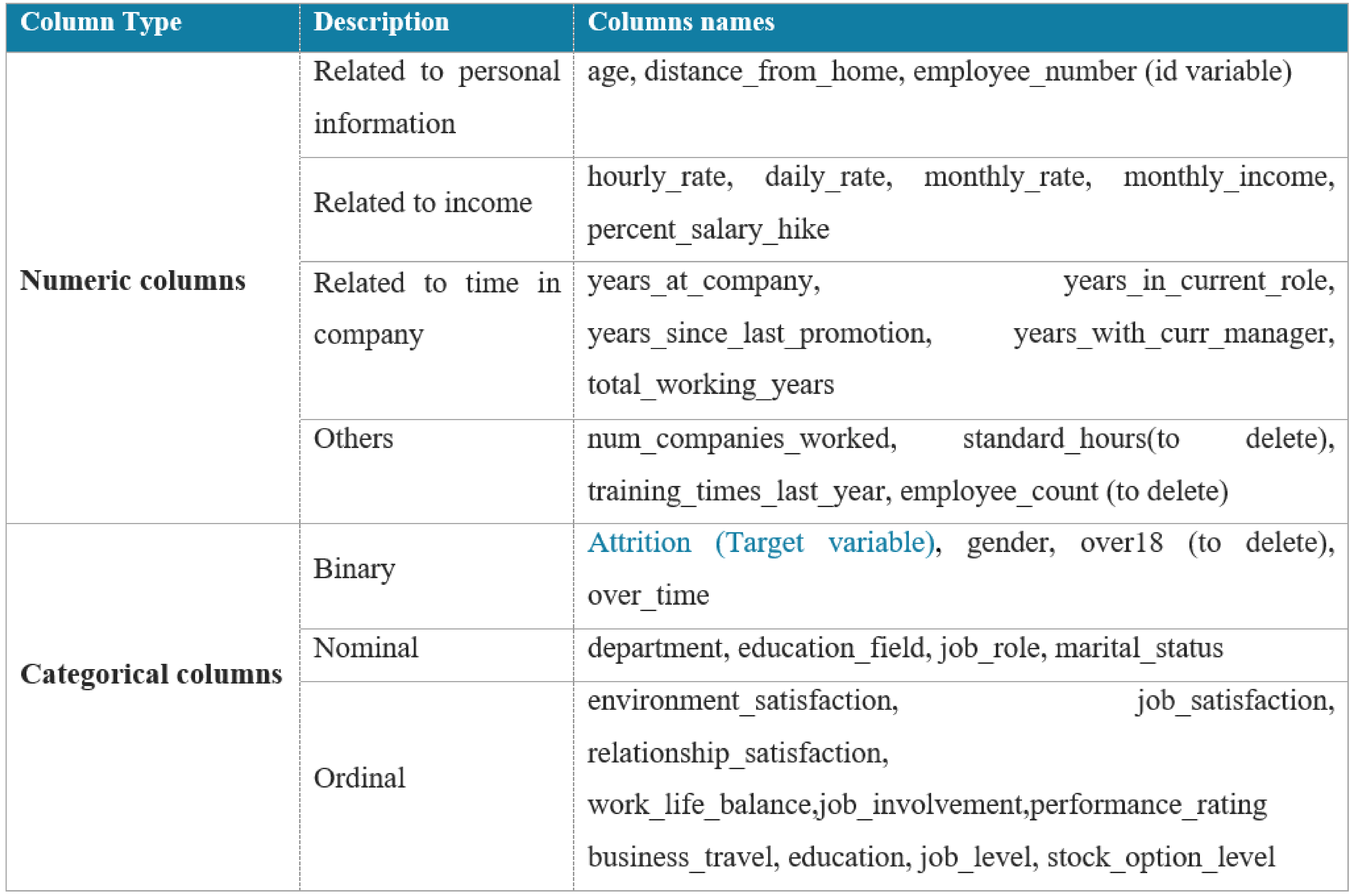
## SOLUTION

Build model that Predicts the appropriate reasons that led to the attrition decision in order to reduce the attrition within a company  in the future

# *Data Description*

- Obtained from kaggle.com
- Fictional dataset created by IBM data scientist
- 1470 rows, 35 columns

# Data Description



| Column Type | Description | Columns names |
|---|---|---|
| **Numeric columns** | Related to personal information | age, distance_from_home, employee_number (id variable) |
| | Related to income | hourly_rate, daily_rate, monthly_rate, monthly_income, percent_salary_hike |
| | Related to time in company | years_at_company, years_in_current_role, years_since_last_promotion, years_with_curr_manager, total_working_years |
| | Others | num_companies_worked, standard_hours(to delete), training_times_last_year, employee_count (to delete) |
| **Categorical columns** | Binary | Attrition (Target variable), gender, over18 (to delete), over_time |
| | Nominal | department, education_field, job_role, marital_status |
| | Ordinal | environment_satisfaction, job_satisfaction, relationship_satisfaction, work_life_balance,job_involvement,performance_rating business_travel, education, job_level, stock_option_level |

# Pre processing steps

EDA

Feature engineering

Scaling

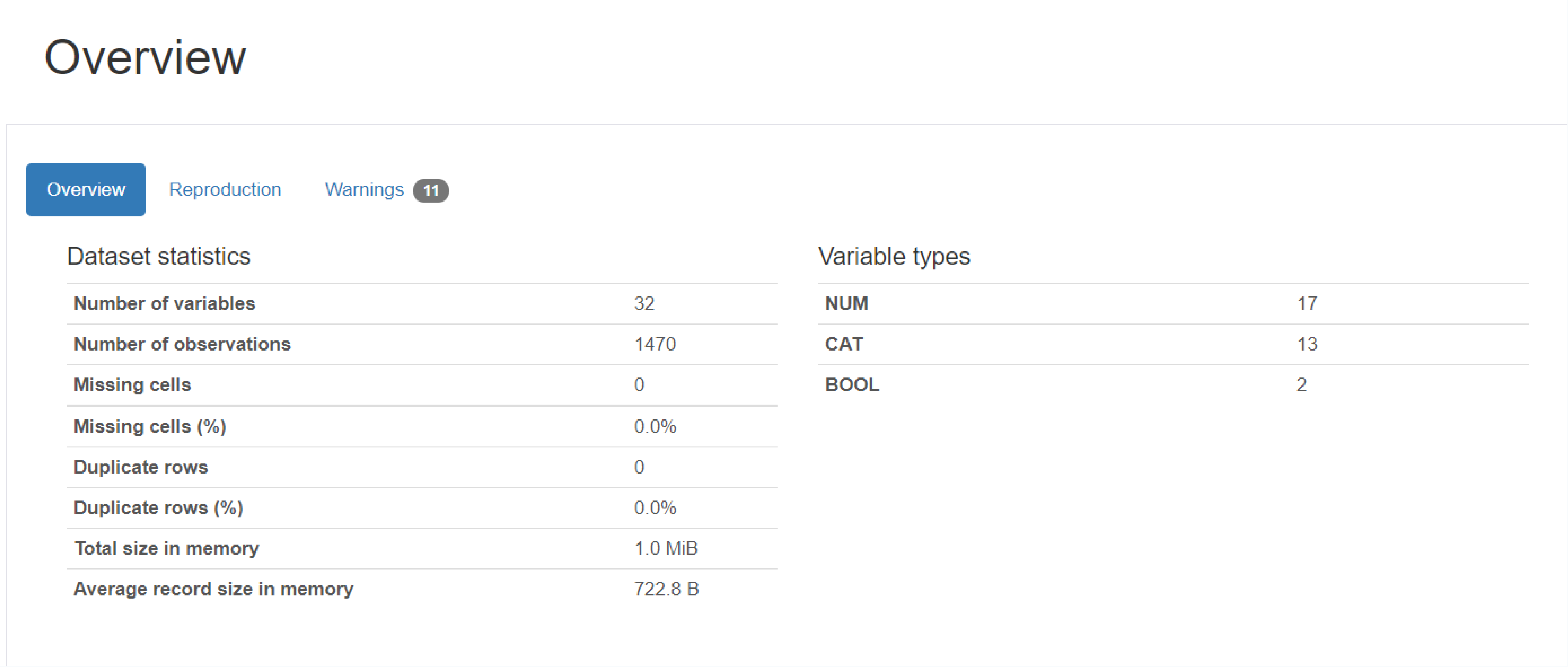
Sampling

EDA

## INSIGHT

Part of the analysis using pandas profiling shows  overview of the dataset

# Using pandas profiling



Overview

Dataset statistics

| | |
|---|---|
| Number of variables | 32 |
| Number of observations | 1470 |
| Missing cells | 0 |
| Missing cells (%) | 0.0% |
| Duplicate rows | 0 |
| Duplicate rows (%) | 0.0% |
| Total size in memory | 1.0 MiB |
| Average record size in memory | 722.8 B |

Variable types

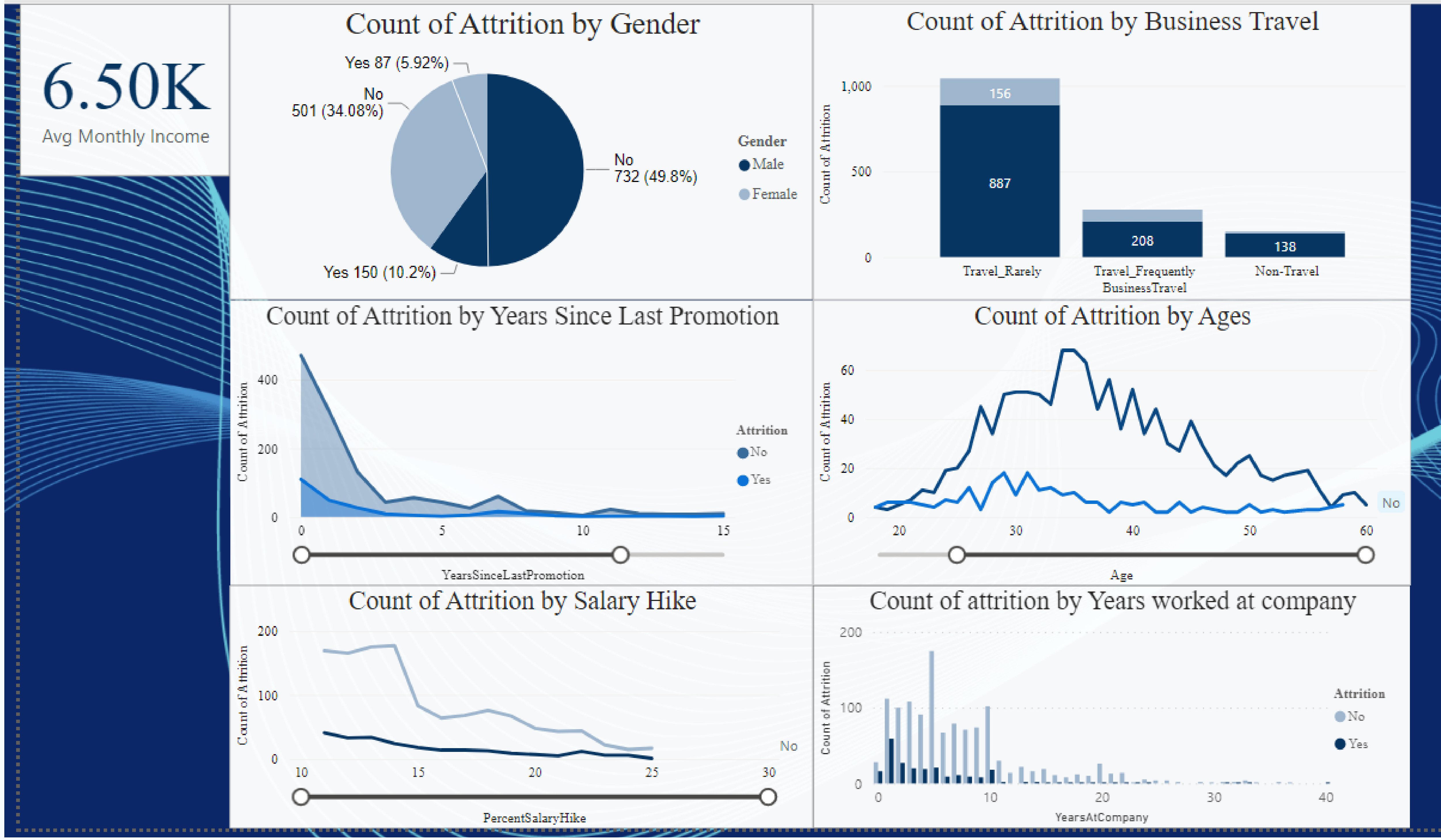
| | |
|---|---|
| NUM | 17 |
| CAT | 13 |
| BOOL | 2 |

## EDA

# Using Power BI

### INSIGHT

Exploring the relations between Attrition and other features in order to have future look to the modeling process.

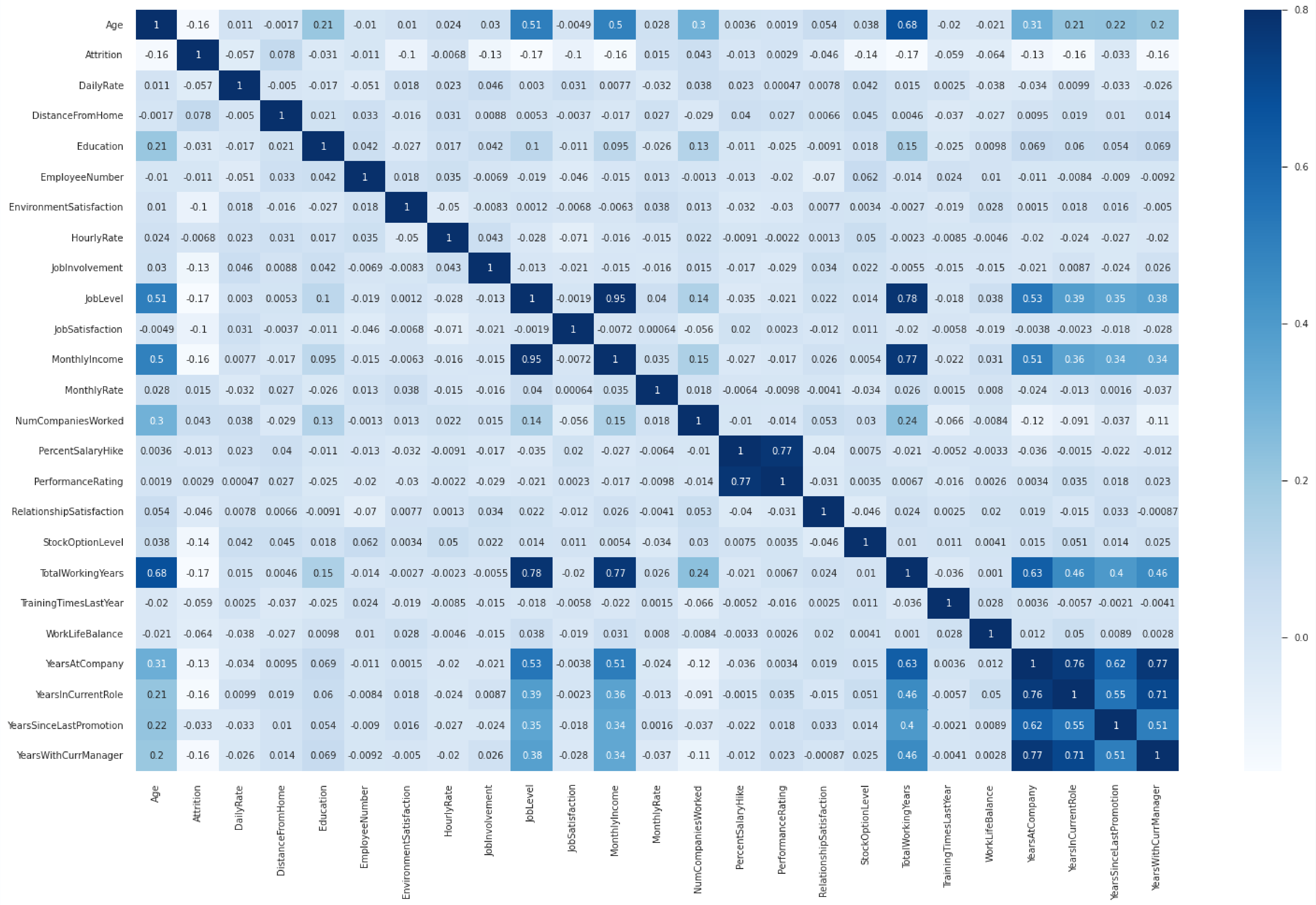


# EDA

# Correlations

## INSIGHT

No segnificant corr between attrition and any feature



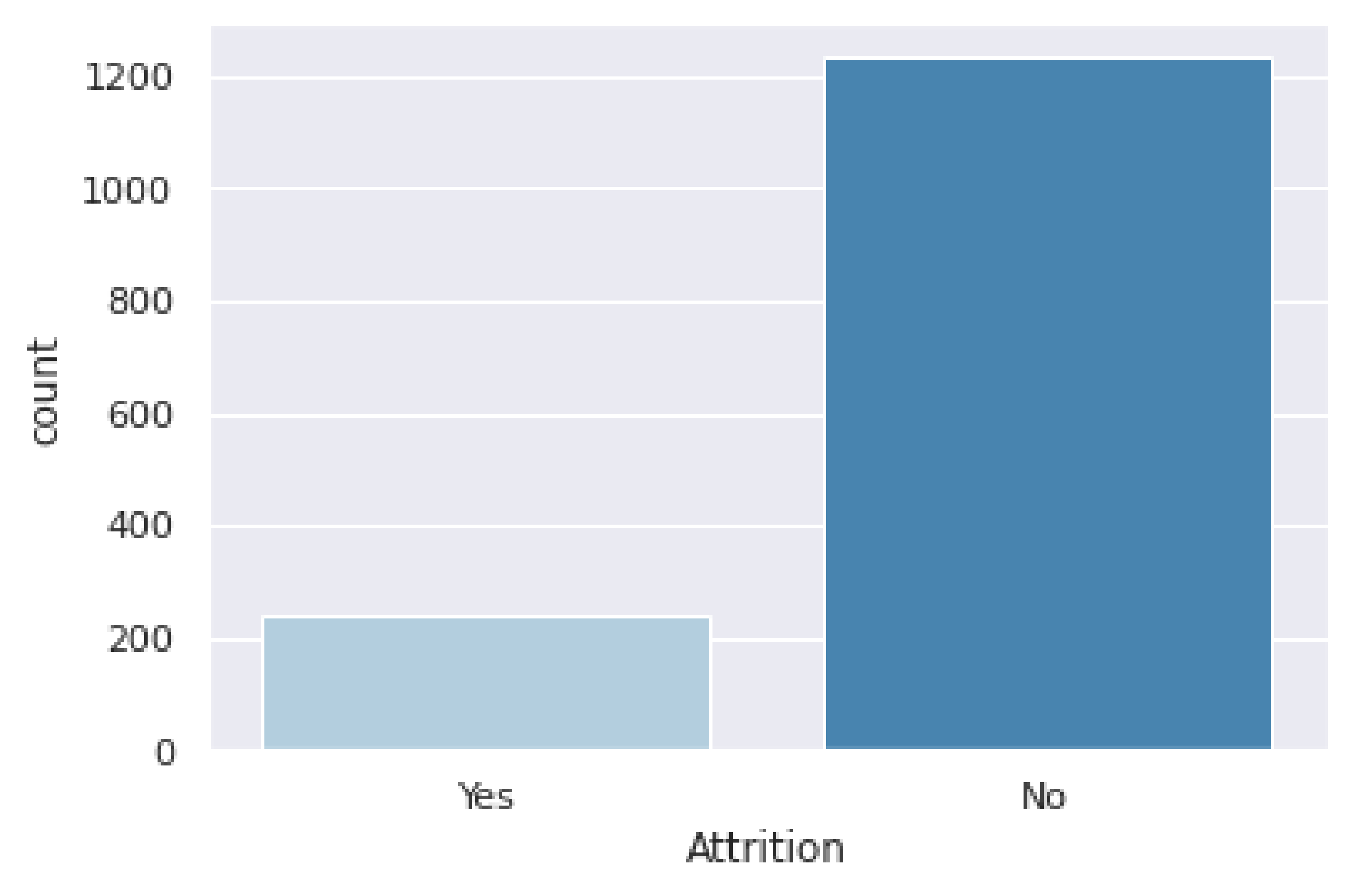
| | Age | Attrition | DailyRate | DistanceFromHome | Education | EmployeeNumber | EnvironmentSatisfaction | HourlyRate | JobInvolvement | JobLevel | JobSatisfaction | MonthlyIncome | MonthlyRate | NumCompaniesWorked | PercentSalaryHike | PerformanceRating | RelationshipSatisfaction | StockOptionLevel | TotalWorkingYears | TrainingTimesLastYear | WorkLifeBalance | YearsAtCompany | YearsInCurrentRole | YearsSinceLastPromotion | YearsWithCurrManager |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | 1 | -0.16 | 0.011 | -0.0017 | 0.21 | -0.01 | 0.01 | 0.024 | 0.03 | 0.51 | -0.0049 | 0.5 | 0.028 | 0.3 | 0.0036 | 0.0019 | 0.054 | 0.038 | 0.68 | -0.02 | -0.021 | 0.31 | 0.21 | 0.22 | 0.2 |
| Attrition | -0.16 | 1 | -0.057 | 0.078 | -0.031 | -0.011 | -0.1 | -0.0068 | -0.13 | -0.17 | -0.1 | -0.16 | 0.015 | 0.043 | -0.013 | 0.0029 | -0.046 | -0.14 | -0.17 | -0.059 | -0.064 | -0.13 | -0.16 | -0.033 | -0.16 |
| DailyRate | 0.011 | -0.057 | 1 | -0.005 | -0.017 | -0.051 | 0.018 | 0.023 | 0.046 | 0.003 | 0.031 | 0.0077 | -0.032 | 0.038 | 0.023 | 0.00047 | 0.0078 | 0.042 | 0.015 | 0.0025 | -0.038 | -0.034 | 0.0099 | -0.033 | -0.026 |
| DistanceFromHome | -0.0017 | 0.078 | -0.005 | 1 | 0.021 | 0.033 | -0.016 | 0.031 | 0.0088 | 0.0053 | -0.0037 | -0.017 | 0.027 | -0.029 | 0.04 | 0.027 | 0.0066 | 0.045 | 0.0046 | -0.037 | -0.027 | 0.0095 | 0.019 | 0.01 | 0.014 |
| Education | 0.21 | -0.031 | -0.017 | 0.021 | 1 | 0.042 | -0.027 | 0.017 | 0.042 | 0.1 | -0.011 | 0.095 | -0.026 | 0.13 | -0.011 | -0.025 | -0.0091 | 0.018 | 0.15 | -0.025 | 0.0098 | 0.069 | 0.06 | 0.054 | 0.069 |
| EmployeeNumber | -0.01 | -0.011 | -0.051 | 0.033 | 0.042 | 1 | 0.018 | 0.035 | -0.0069 | -0.019 | -0.046 | -0.015 | 0.013 | -0.0013 | -0.013 | -0.02 | -0.07 | 0.062 | -0.014 | 0.024 | 0.01 | -0.011 | -0.0084 | -0.009 | -0.0092 |
| EnvironmentSatisfaction | 0.01 | -0.1 | 0.018 | -0.016 | -0.027 | 0.018 | 1 | -0.05 | -0.0083 | 0.0012 | -0.0068 | -0.0063 | 0.038 | 0.013 | -0.032 | -0.03 | 0.0077 | 0.0034 | -0.0027 | -0.019 | 0.028 | 0.0015 | 0.018 | 0.016 | -0.005 |
| HourlyRate | 0.024 | -0.0068 | 0.023 | 0.031 | 0.017 | 0.035 | -0.05 | 1 | 0.043 | -0.028 | -0.071 | -0.016 | -0.015 | 0.022 | -0.0091 | -0.0022 | 0.0013 | 0.05 | -0.0023 | -0.0085 | -0.0046 | -0.02 | -0.024 | -0.027 | -0.02 |
| JobInvolvement | 0.03 | -0.13 | 0.046 | 0.0088 | 0.042 | -0.0069 | -0.0083 | 0.043 | 1 | -0.013 | -0.021 | -0.015 | -0.016 | 0.015 | -0.017 | -0.029 | 0.034 | 0.022 | -0.0055 | -0.015 | -0.015 | -0.021 | 0.0087 | -0.024 | 0.026 |
| JobLevel | 0.51 | -0.17 | 0.003 | 0.0053 | 0.1 | -0.019 | 0.0012 | -0.028 | -0.013 | 1 | -0.0019 | 0.95 | 0.04 | 0.14 | -0.035 | -0.021 | 0.022 | 0.014 | 0.78 | -0.018 | 0.038 | 0.53 | 0.39 | 0.35 | 0.38 |
| JobSatisfaction | -0.0049 | -0.1 | 0.031 | -0.0037 | -0.011 | -0.046 | -0.0068 | -0.071 | -0.021 | -0.0019 | 1 | -0.0072 | 0.00064 | -0.056 | 0.02 | 0.0023 | -0.012 | 0.011 | -0.02 | -0.0058 | -0.019 | -0.0038 | -0.0023 | -0.018 | -0.028 |
| MonthlyIncome | 0.5 | -0.16 | 0.0077 | -0.017 | 0.095 | -0.015 | -0.0063 | -0.016 | -0.015 | 0.95 | -0.0072 | 1 | 0.035 | 0.15 | -0.027 | -0.017 | 0.026 | 0.0054 | 0.77 | -0.022 | 0.031 | 0.51 | 0.36 | 0.34 | 0.34 |
| MonthlyRate | 0.028 | 0.015 | -0.032 | 0.027 | -0.026 | 0.013 | 0.038 | -0.015 | -0.016 | 0.04 | 0.00064 | 0.035 | 1 | 0.018 | -0.0064 | -0.0098 | -0.0041 | -0.034 | 0.026 | 0.0015 | 0.008 | -0.024 | -0.013 | 0.0016 | -0.037 |
| NumCompaniesWorked | 0.3 | 0.043 | 0.038 | -0.029 | 0.13 | -0.0013 | 0.013 | 0.022 | 0.015 | 0.14 | -0.056 | 0.15 | 0.018 | 1 | -0.01 | -0.014 | 0.053 | 0.03 | 0.24 | -0.066 | -0.0084 | -0.12 | -0.091 | -0.037 | -0.11 |
| PercentSalaryHike | 0.0036 | -0.013 | 0.023 | 0.04 | -0.011 | -0.013 | -0.032 | -0.0091 | -0.017 | -0.035 | 0.02 | -0.027 | -0.0064 | -0.01 | 1 | 0.77 | -0.04 | 0.0075 | -0.021 | -0.0052 | -0.0033 | -0.036 | -0.0015 | -0.022 | -0.012 |
| PerformanceRating | 0.0019 | 0.0029 | 0.00047 | 0.027 | -0.025 | -0.02 | -0.03 | -0.0022 | -0.029 | -0.021 | 0.0023 | -0.017 | -0.0098 | -0.014 | 0.77 | 1 | -0.031 | 0.0035 | 0.0067 | -0.016 | 0.0026 | 0.0034 | 0.035 | 0.018 | 0.023 |
| RelationshipSatisfaction | 0.054 | -0.046 | 0.0078 | 0.0066 | -0.0091 | -0.07 | 0.0077 | 0.0013 | 0.034 | 0.022 | -0.012 | 0.026 | -0.0041 | 0.053 | -0.04 | -0.031 | 1 | -0.046 | 0.024 | 0.0025 | 0.02 | 0.019 | -0.015 | 0.033 | -0.00087 |
| StockOptionLevel | 0.038 | -0.14 | 0.042 | 0.045 | 0.018 | 0.062 | 0.0034 | 0.05 | 0.022 | 0.014 | 0.011 | 0.0054 | -0.034 | 0.03 | 0.0075 | 0.0035 | -0.046 | 1 | 0.01 | 0.011 | 0.0041 | 0.015 | 0.051 | 0.014 | 0.025 |
| TotalWorkingYears | 0.68 | -0.17 | 0.015 | 0.0046 | 0.15 | -0.014 | -0.0027 | -0.0023 | -0.0055 | 0.78 | -0.02 | 0.77 | 0.026 | 0.24 | -0.021 | 0.0067 | 0.024 | 0.01 | 1 | -0.036 | 0.001 | 0.63 | 0.46 | 0.4 | 0.46 |
| TrainingTimesLastYear | -0.02 | -0.059 | 0.0025 | -0.037 | -0.025 | 0.024 | -0.019 | -0.0085 | -0.015 | -0.018 | -0.0058 | -0.022 | 0.0015 | -0.066 | -0.0052 | -0.016 | 0.0025 | 0.011 | -0.036 | 1 | 0.028 | 0.0036 | -0.0057 | -0.0021 | -0.0041 |
| WorkLifeBalance | -0.021 | -0.064 | -0.038 | -0.027 | 0.0098 | 0.01 | 0.028 | -0.0046 | -0.015 | 0.038 | -0.019 | 0.031 | 0.008 | -0.0084 | -0.0033 | 0.0026 | 0.02 | 0.0041 | 0.001 | 0.028 | 1 | 0.012 | 0.05 | 0.0089 | 0.0028 |
| YearsAtCompany | 0.31 | -0.13 | -0.034 | 0.0095 | 0.069 | -0.011 | 0.0015 | -0.02 | -0.021 | 0.53 | -0.0038 | 0.51 | -0.024 | -0.12 | -0.036 | 0.0034 | 0.019 | 0.015 | 0.63 | 0.0036 | 0.012 | 1 | 0.76 | 0.62 | 0.77 |
| YearsInCurrentRole | 0.21 | -0.16 | 0.0099 | 0.019 | 0.06 | -0.0084 | 0.018 | -0.024 | 0.0087 | 0.39 | -0.0023 | 0.36 | -0.013 | -0.091 | -0.0015 | 0.035 | -0.015 | 0.051 | 0.46 | -0.0057 | 0.05 | 0.76 | 1 | 0.55 | 0.71 |
| YearsSinceLastPromotion | 0.22 | -0.033 | -0.033 | 0.01 | 0.054 | -0.009 | 0.016 | -0.027 | -0.024 | 0.35 | -0.018 | 0.34 | 0.0016 | -0.037 | -0.022 | 0.018 | 0.033 | 0.014 | 0.4 | -0.0021 | 0.0089 | 0.62 | 0.55 | 1 | 0.51 |
| YearsWithCurrManager | 0.2 | -0.16 | -0.026 | 0.014 | 0.069 | -0.0092 | -0.005 | -0.02 | 0.026 | 0.38 | -0.028 | 0.34 | -0.037 | -0.11 | -0.012 | 0.023 | -0.00087 | 0.025 | 0.46 | -0.0041 | 0.0028 | 0.77 | 0.71 | 0.51 | 1 |

EDA

# Count of Attrition

INSIGHT

Noticing class imbalance!



# *Feature engineering*

- Exclude unnecessary features (over18, standard hours, employee count)
- Convert the attrition column from Yes,No to 1,0
- Split the data into 30% test and 70% train

# Scaling

- Using MinMax scaler

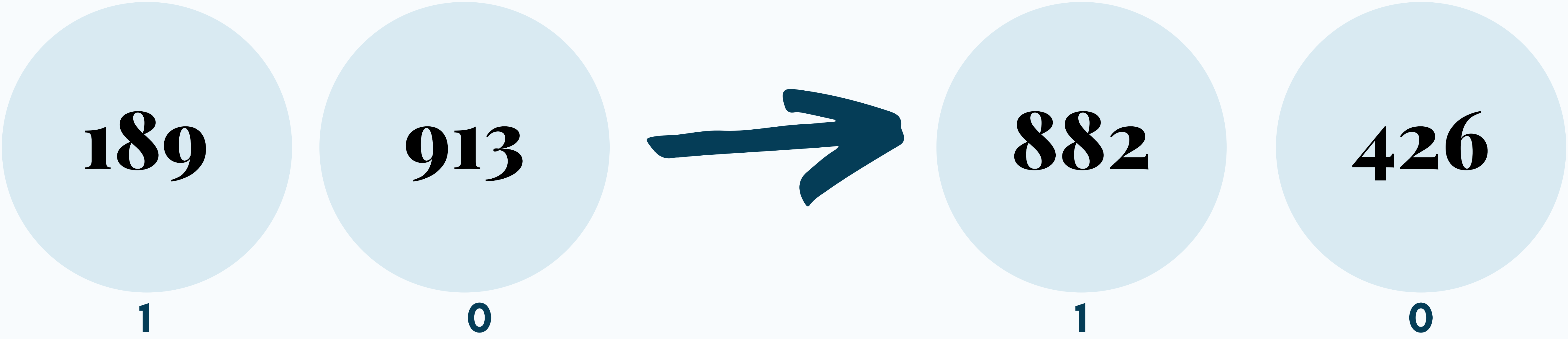
# *Sampling*

## USING HYBRID TECHNIQUE (SMOTE+ENN)

- ENN: algorithm for pattern recognition that predicts the pattern of an unknown test sample hinged on the highest gain of intraclass coherence
- ENN+SMOTE helps in doing extensive data cleaning.
- The misclassification by NN's samples from both the classes are removed.

# *Sampling*

## USING HYBRID TECHNIQUE (SMOTE+ENN)



189

1

913

0

882

1

426

0

# *Modeling and evaluation*

Logistic regression

Decision Tree

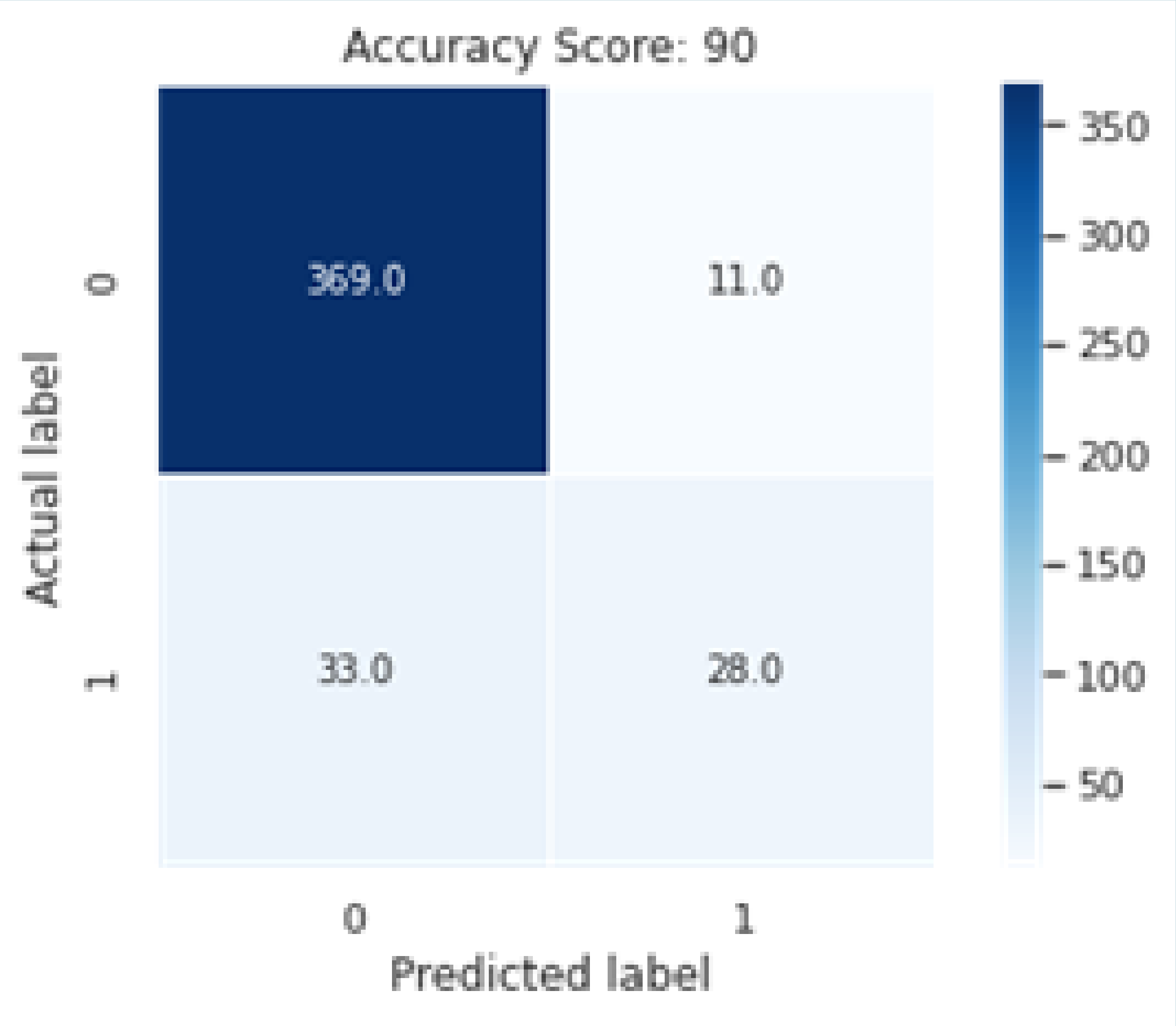
SVM

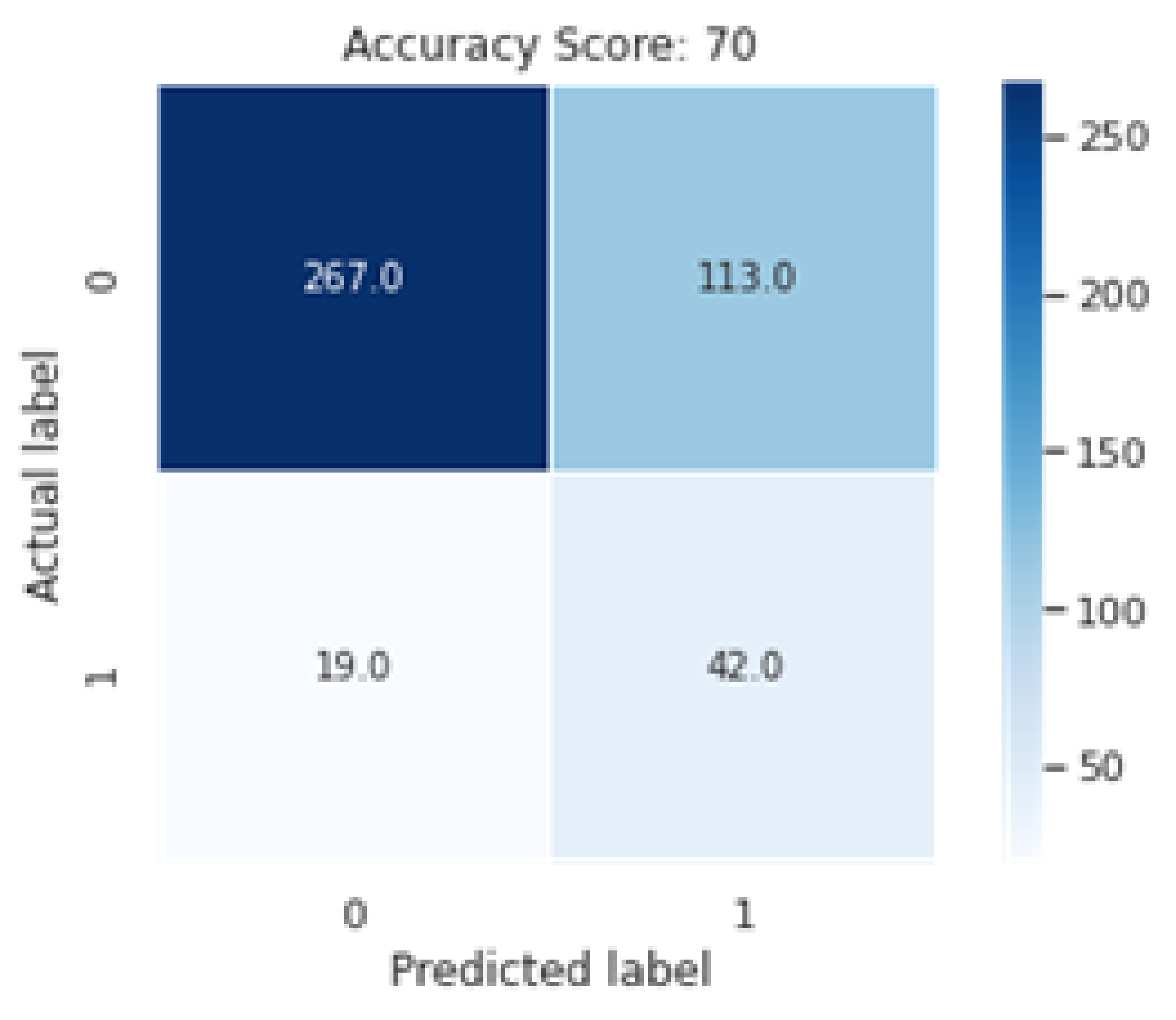
Ensemble: AdaBoost

Models comparison

Optimization

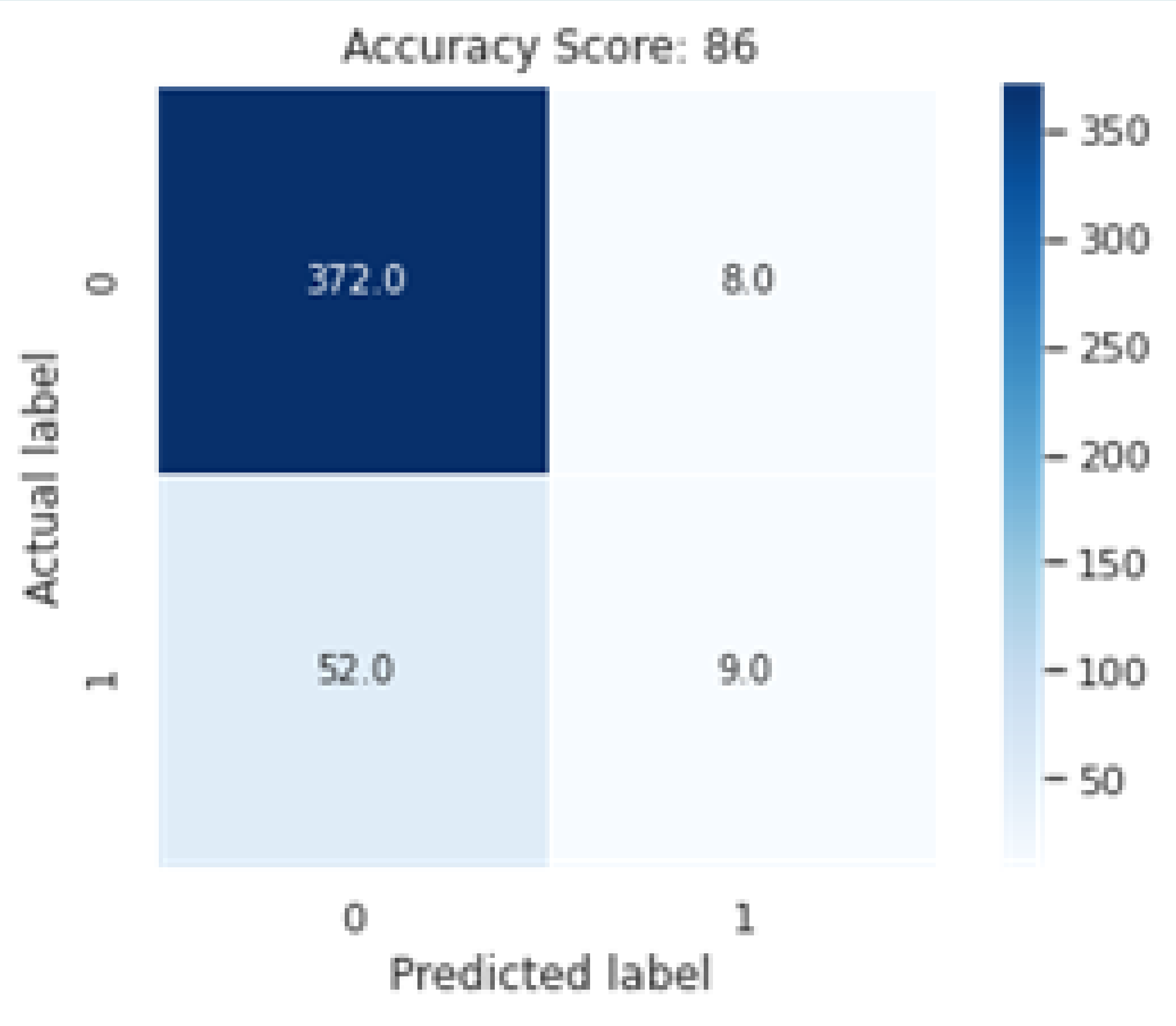
# *Logistic regression*

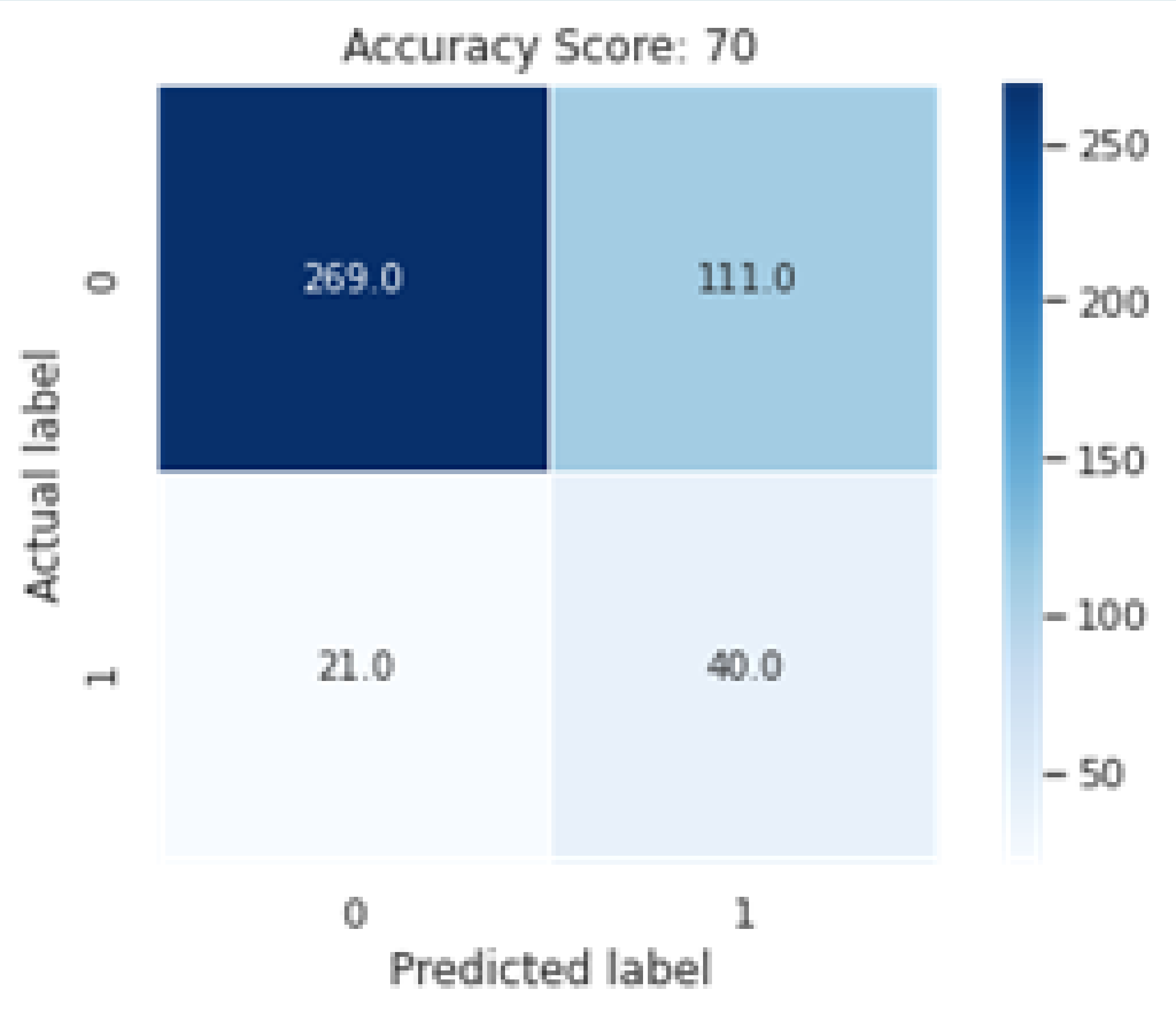




## Confusion matrix before and after sampling

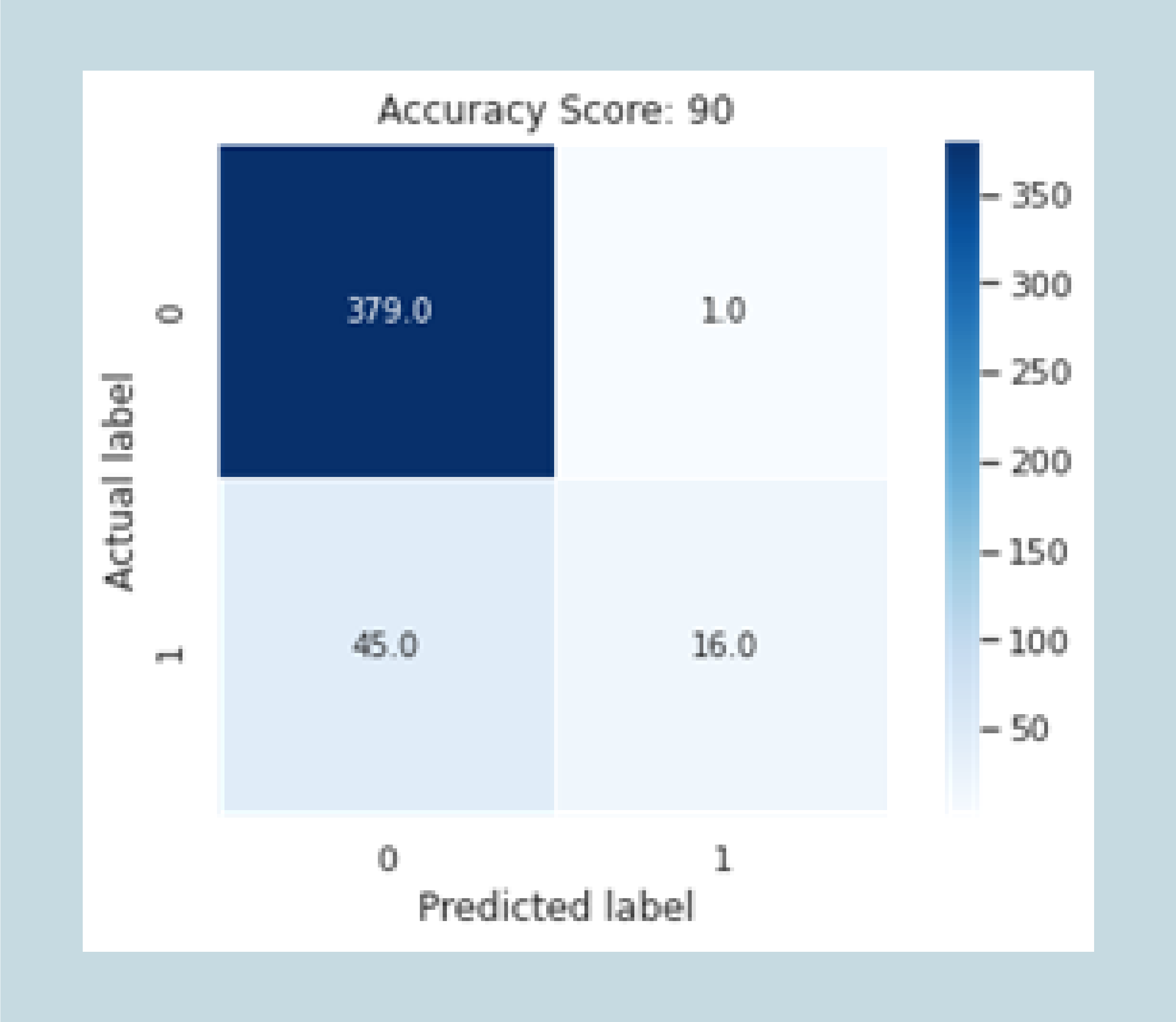
# Decision Tree



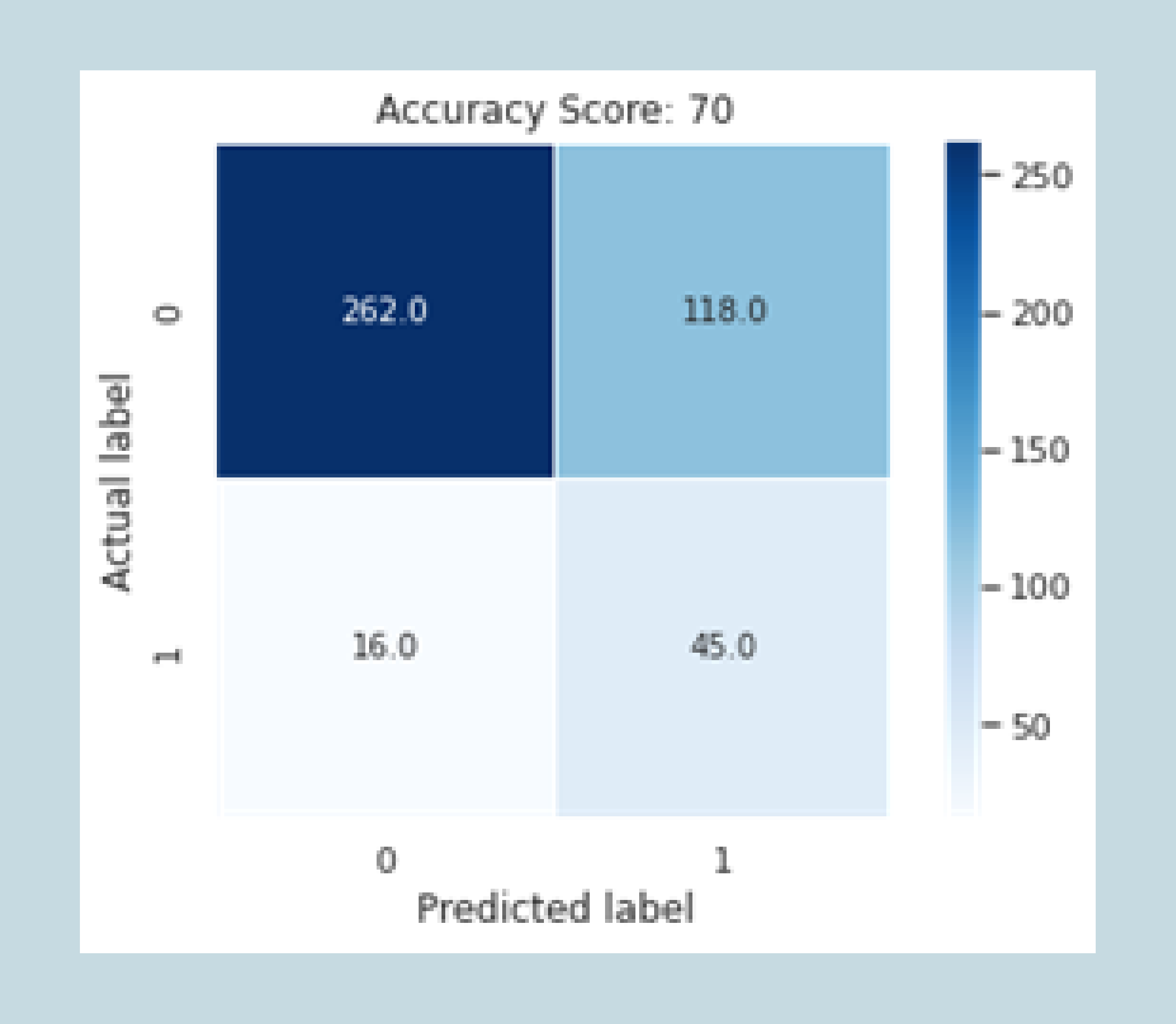


Confusion matrix before and after sampling

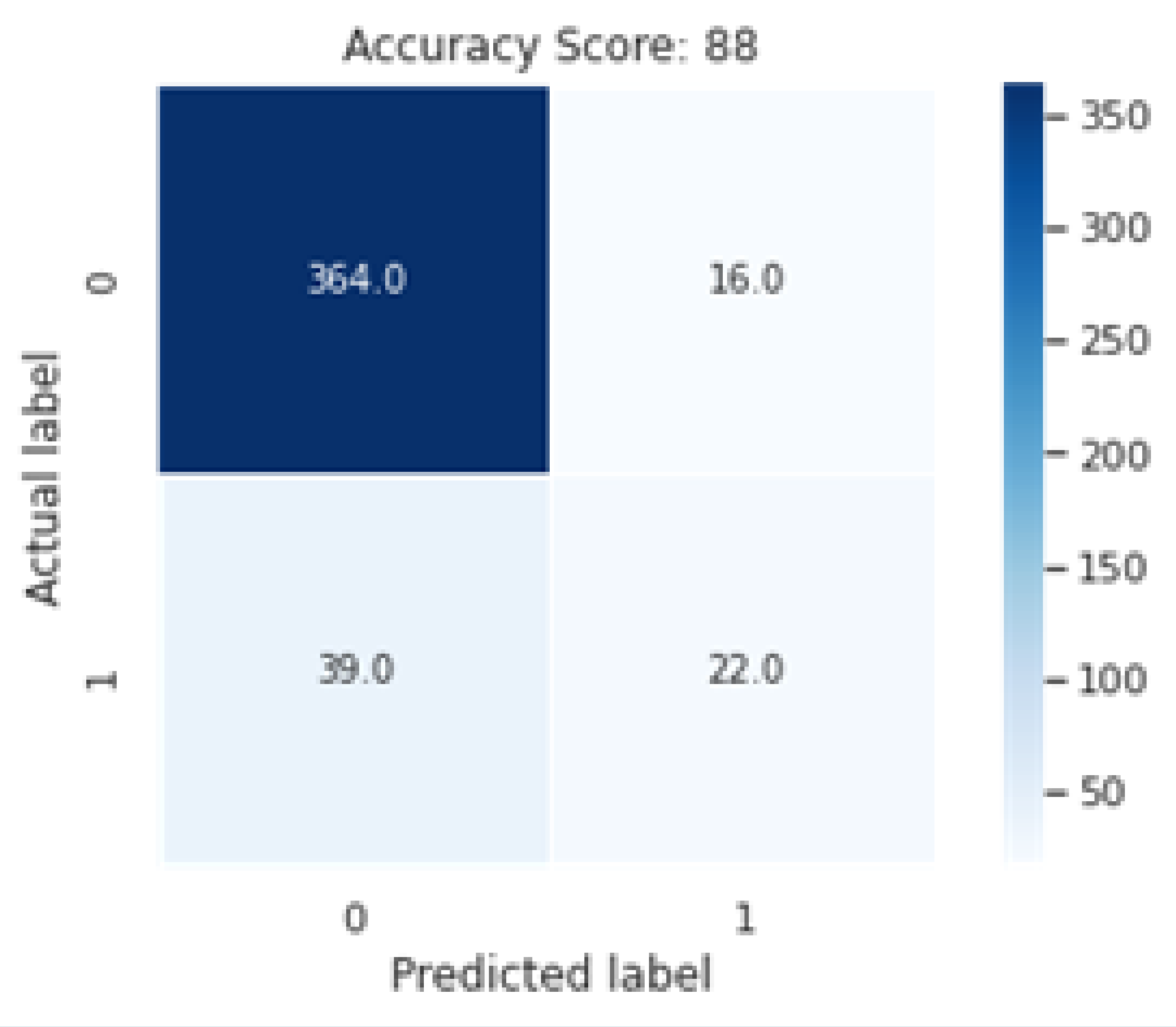
# SVM

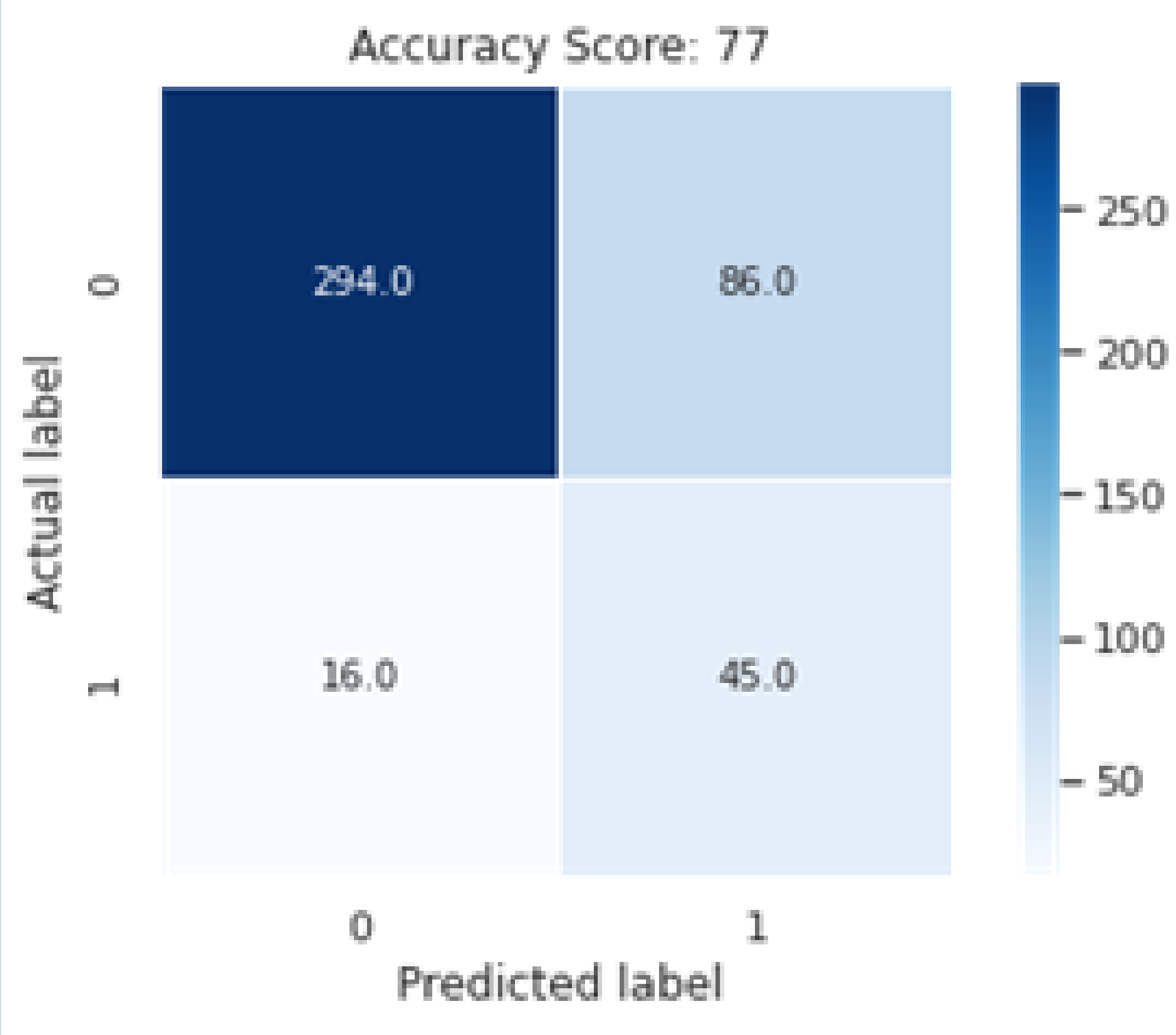


Confusion matrix before and after sampling



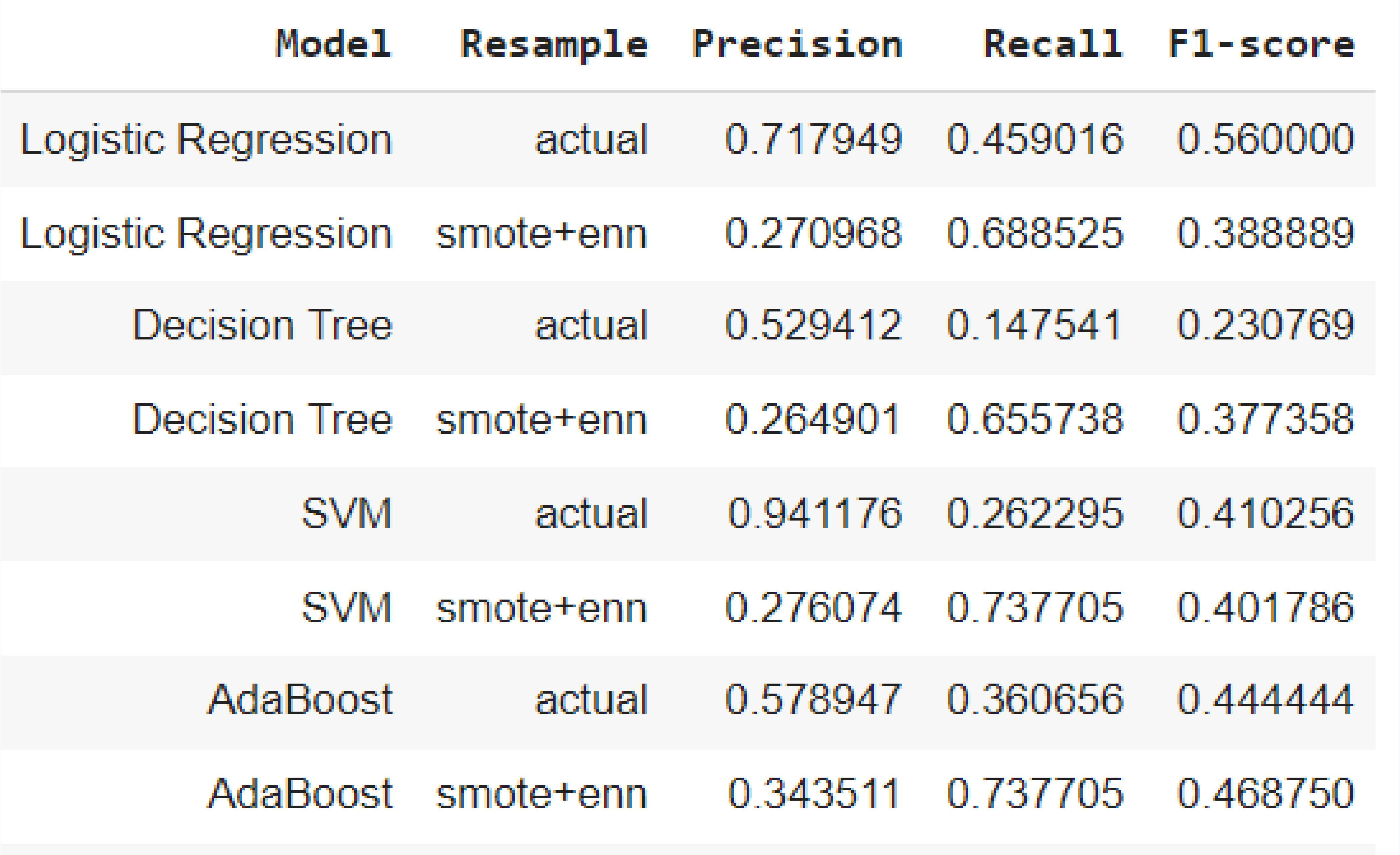
# *AdaBoost*





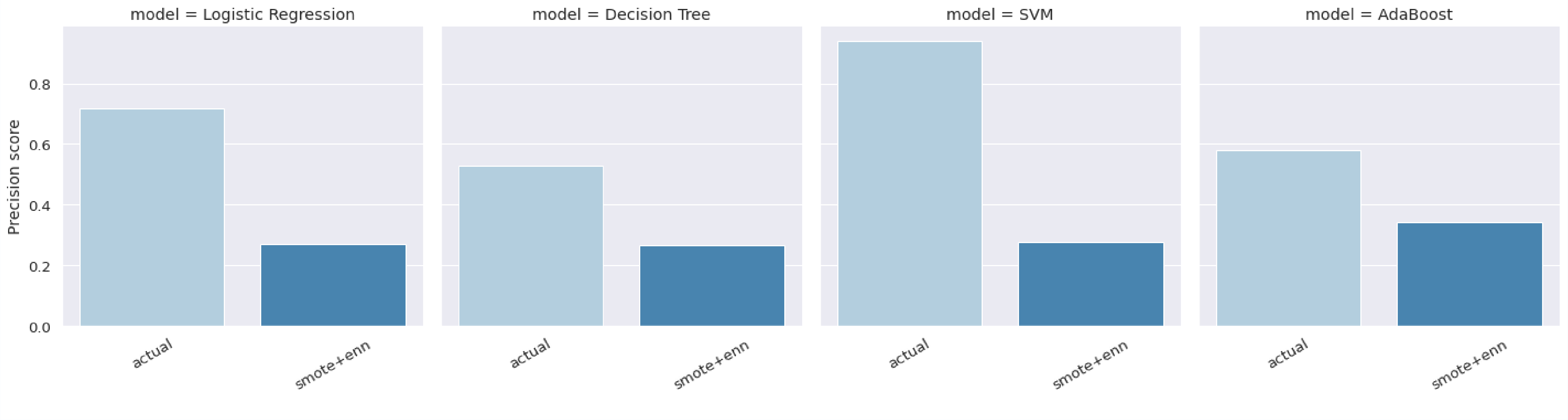
## Confusion matrix before and after sampling

# Models Comparison



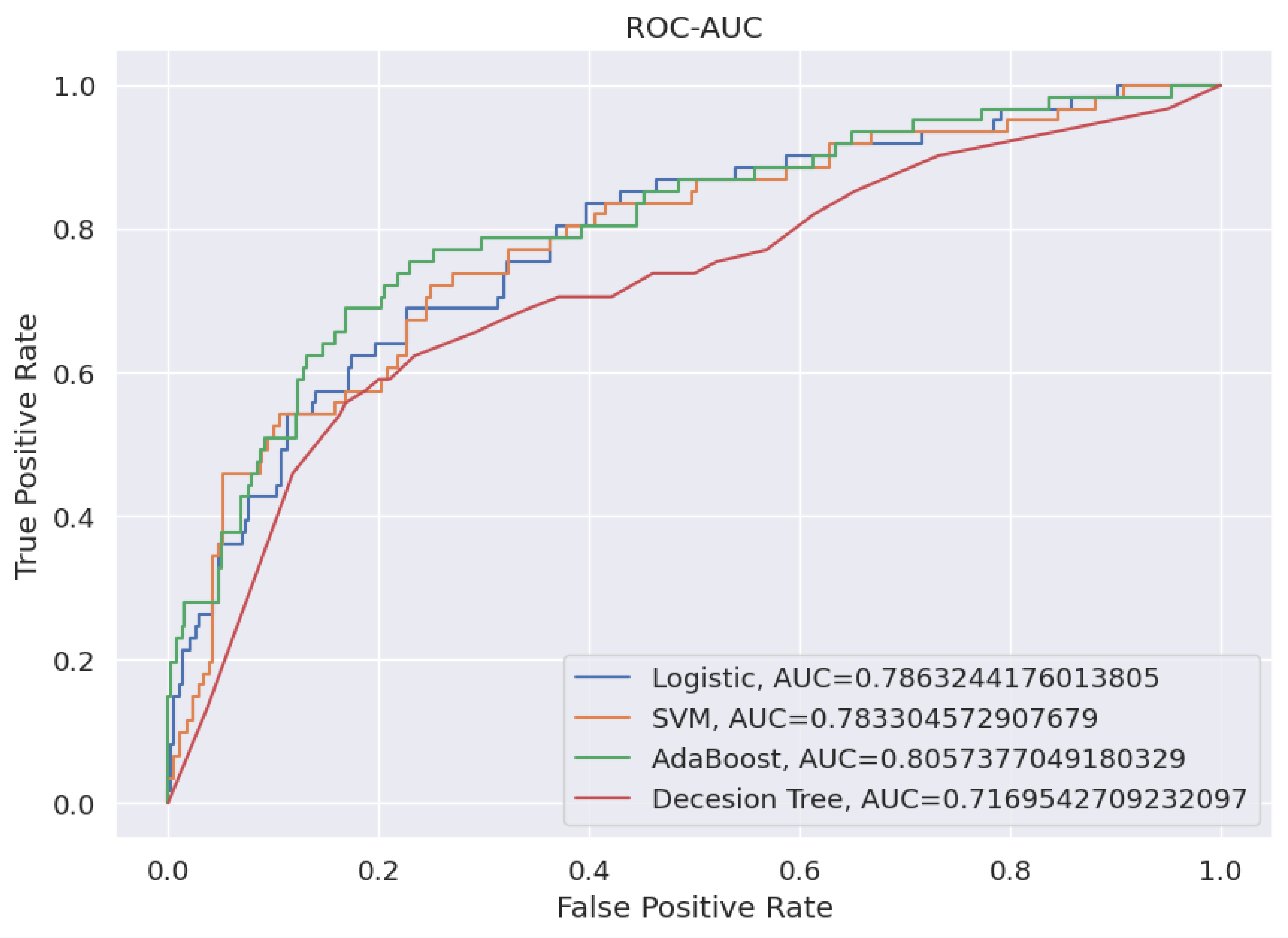
| Model | Resample | Precision | Recall | F1-score |
|---:|---:|---:|---:|---:|
| Logistic Regression | actual | 0.717949 | 0.459016 | 0.560000 |
| Logistic Regression | smote+enn | 0.270968 | 0.688525 | 0.388889 |
| Decision Tree | actual | 0.529412 | 0.147541 | 0.230769 |
| Decision Tree | smote+enn | 0.264901 | 0.655738 | 0.377358 |
| SVM | actual | 0.941176 | 0.262295 | 0.410256 |
| SVM | smote+enn | 0.276074 | 0.737705 | 0.401786 |
| AdaBoost | actual | 0.578947 | 0.360656 | 0.444444 |
| AdaBoost | smote+enn | 0.343511 | 0.737705 | 0.468750 |

# *Models Comparison*



**INSIGHT:** SVM before sampling has best precision score , however we cannot decide yet if it's the best model  or not

# *Models Comparison*

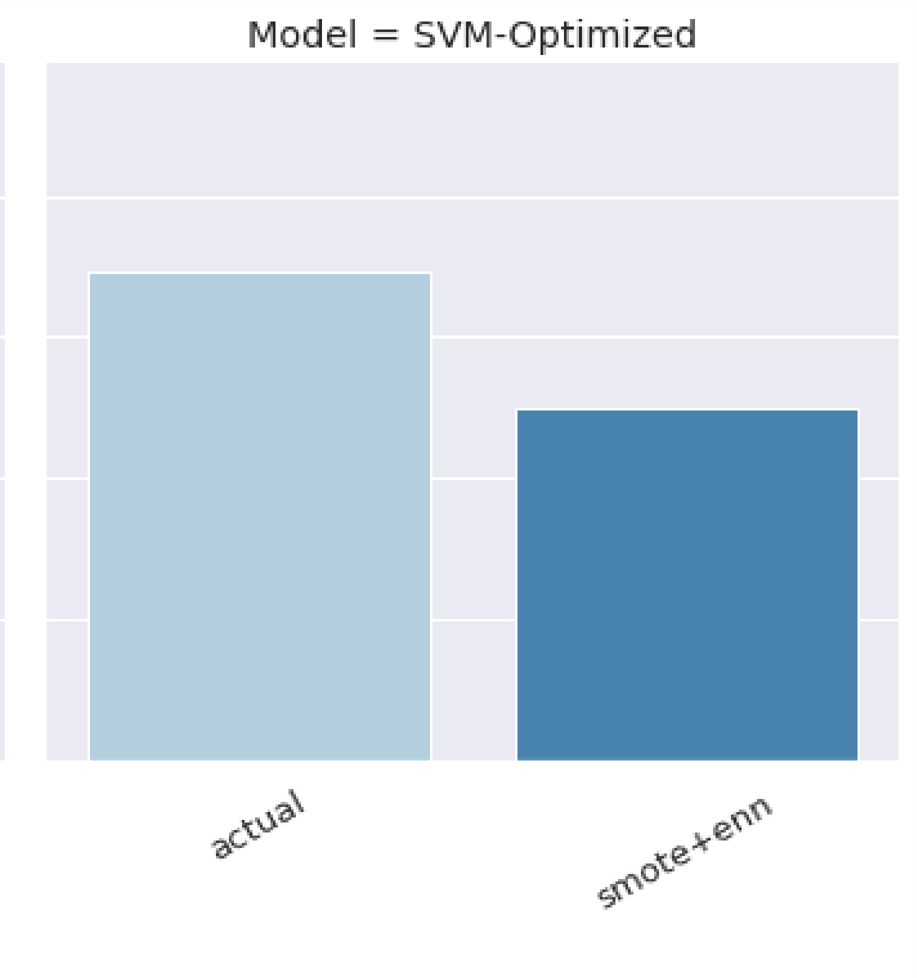


# *Optimization*

## OPTIMIZING SVM

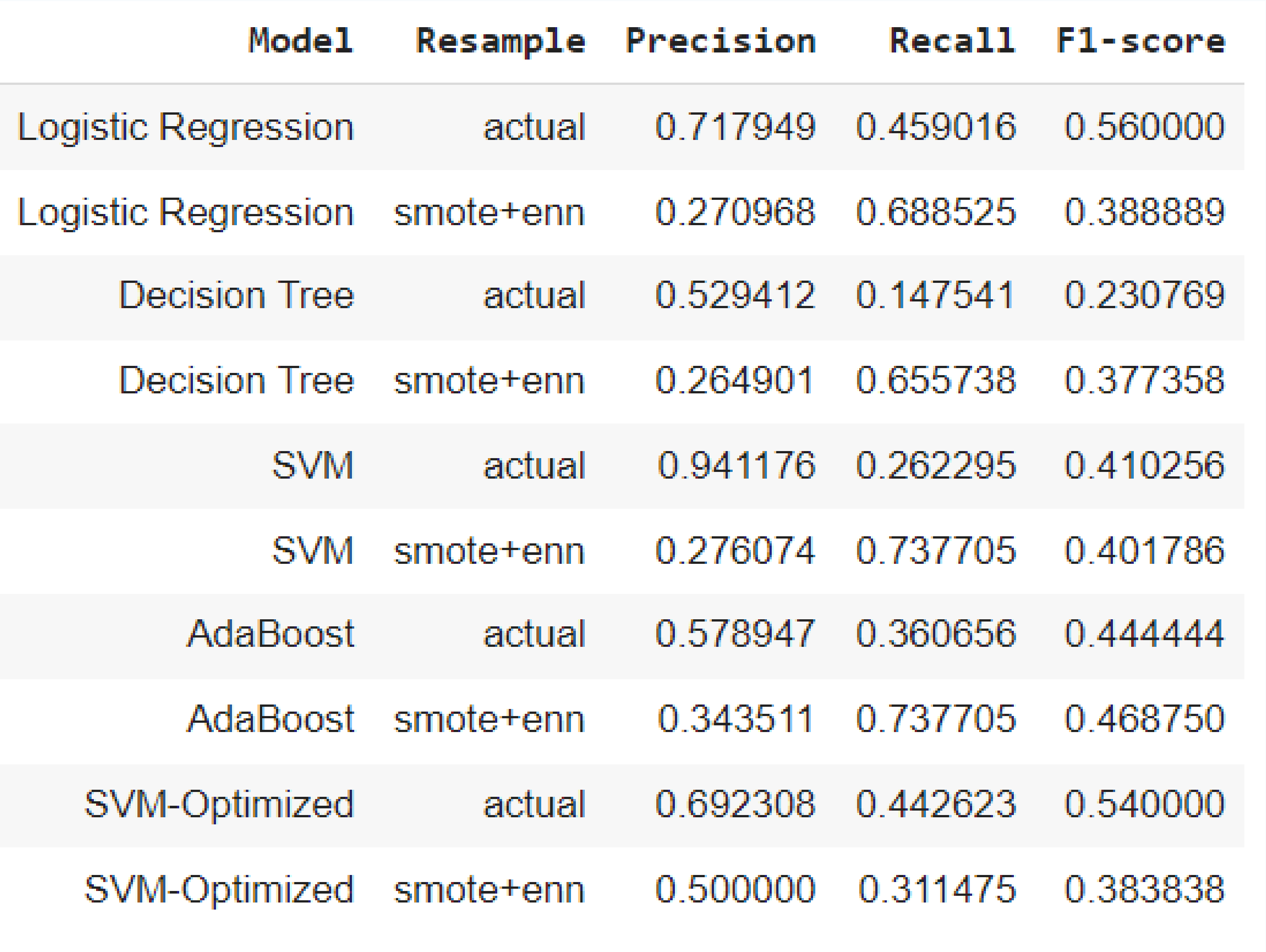
- Using GridSearchCV
- Best parameters were: C=1, Gamma=1, Kernel=RBF

# *Models Comparison: after optimizing*



**INSIGHT:** the precision decreased in actual and increased in sampled

# *Models Comparison: after optimizing*



| Model | Resample | Precision | Recall | F1-score |
|---:|---:|---:|---:|---:|
| Logistic Regression | actual | 0.717949 | 0.459016 | 0.560000 |
| Logistic Regression | smote+enn | 0.270968 | 0.688525 | 0.388889 |
| Decision Tree | actual | 0.529412 | 0.147541 | 0.230769 |
| Decision Tree | smote+enn | 0.264901 | 0.655738 | 0.377358 |
| SVM | actual | 0.941176 | 0.262295 | 0.410256 |
| SVM | smote+enn | 0.276074 | 0.737705 | 0.401786 |
| AdaBoost | actual | 0.578947 | 0.360656 | 0.444444 |
| AdaBoost | smote+enn | 0.343511 | 0.737705 | 0.468750 |
| SVM-Optimized | actual | 0.692308 | 0.442623 | 0.540000 |
| SVM-Optimized | smote+enn | 0.500000 | 0.311475 | 0.383838 |

# Conclusions

- SVM AFTER OPTIMIZATION AND LOGISTIC REGRESSION MAY BE GOOD MODELS FOR OUR CASE
- ADABOOST CLASSIFIER HAS THE LARGEST AUC AMONG ALL MODELS WE CAN SAY THAT ITS ALSO A GOOD CLASSIFIER
- DECESION TREE SEEM TO HAVE BEEN OVERFITTED DUE TO LOW VOLUME DATA IF THE DATA WAS LARGER TREES WOULD PERFORM BETTER

# Thank you!

ANY QUESTIONS?