EMPLOYEE CHURN PREDICTION USING ML MODELS

By AFIYA AFSHEEN

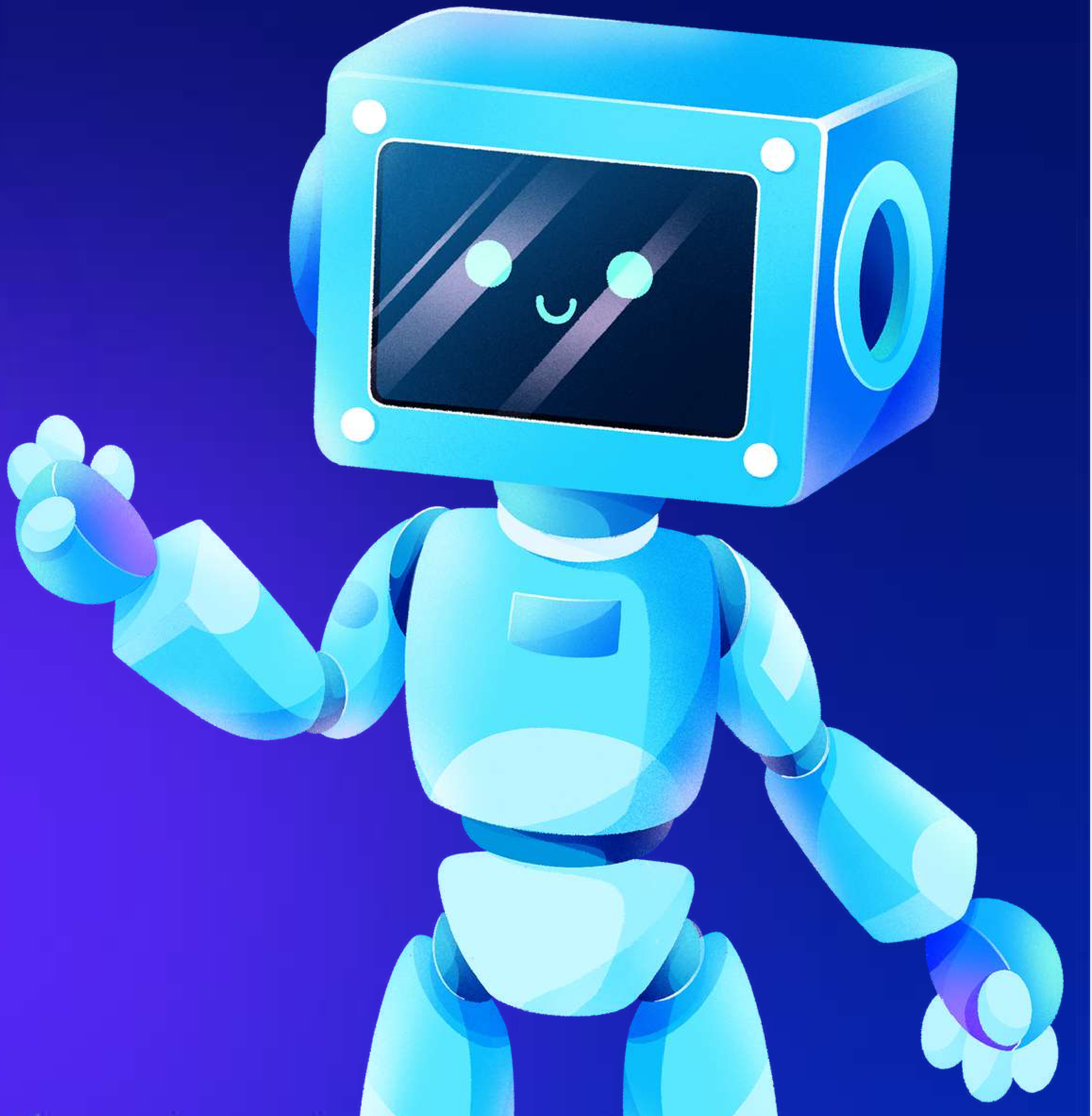
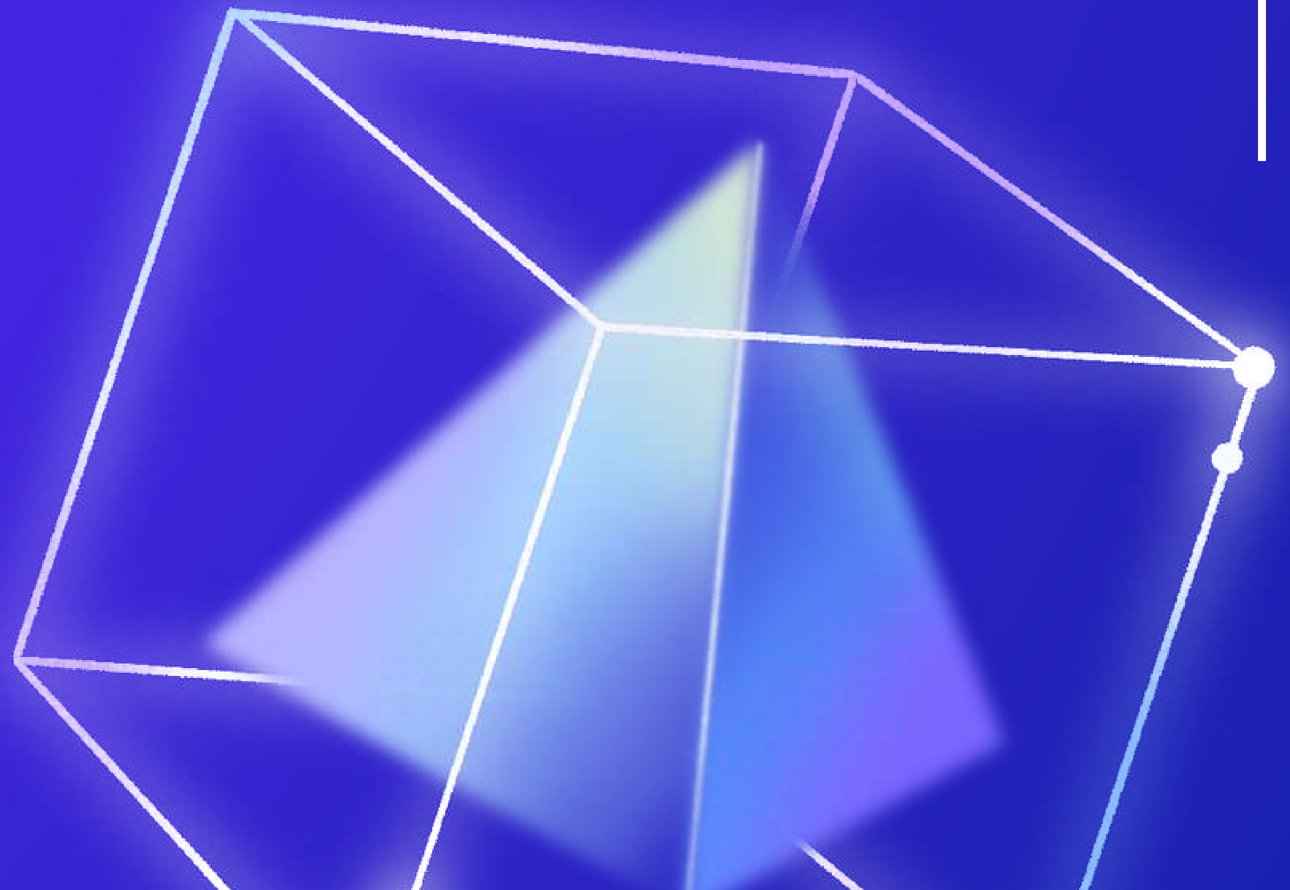
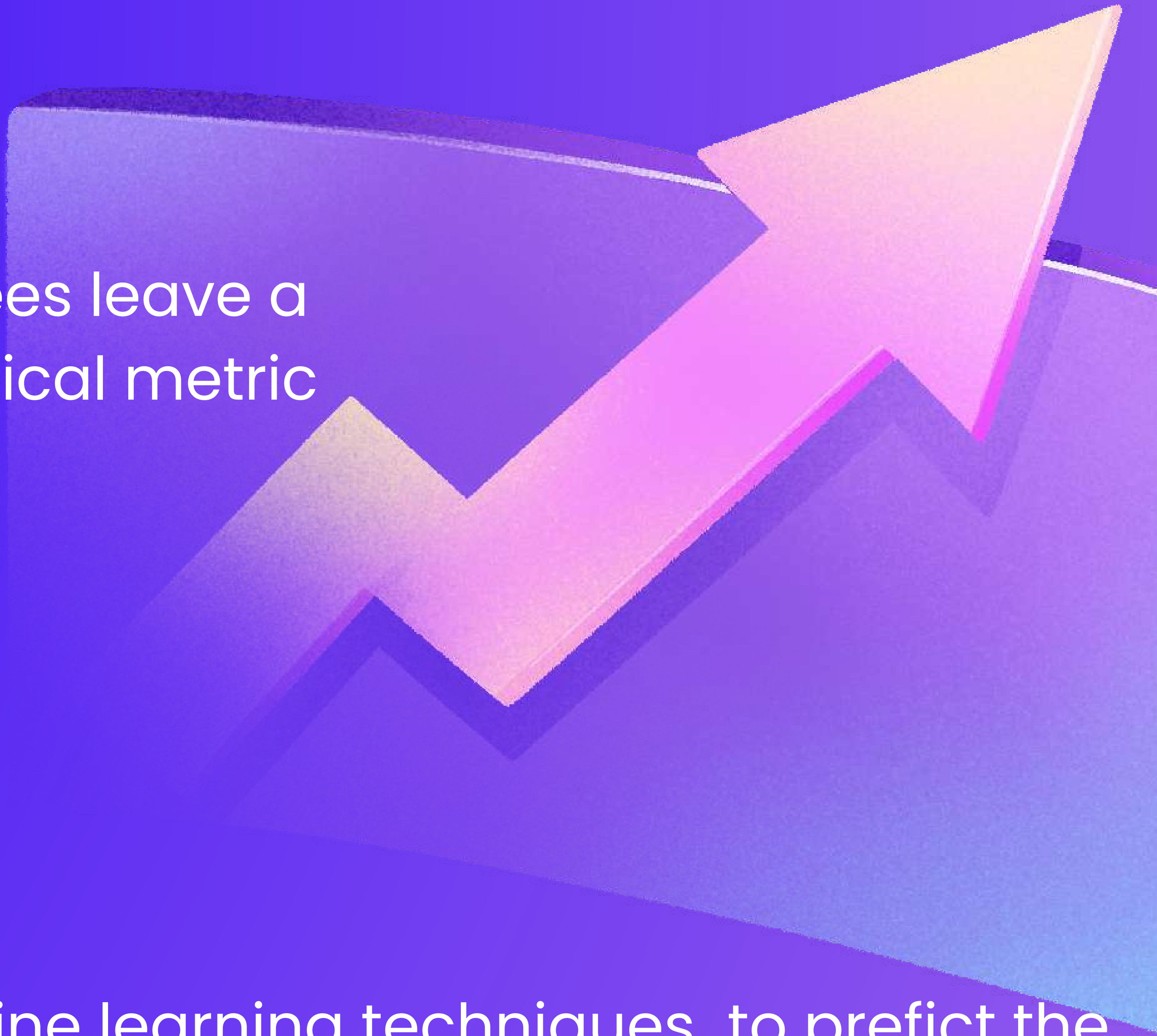


TABLE OF CONTENTS

- Introduction
- Project Flow
- About Dataset
- Basic Cleaning
- Methodology-wise Basic DataFrame
- Visualization
- Correlation Matrix
- Class Imbalnce ,Outliers,Skewness.
- DataFrame_2
- Dataframe_1 (v/s) DataFrame_2
- Feature Selection
- Conclusion



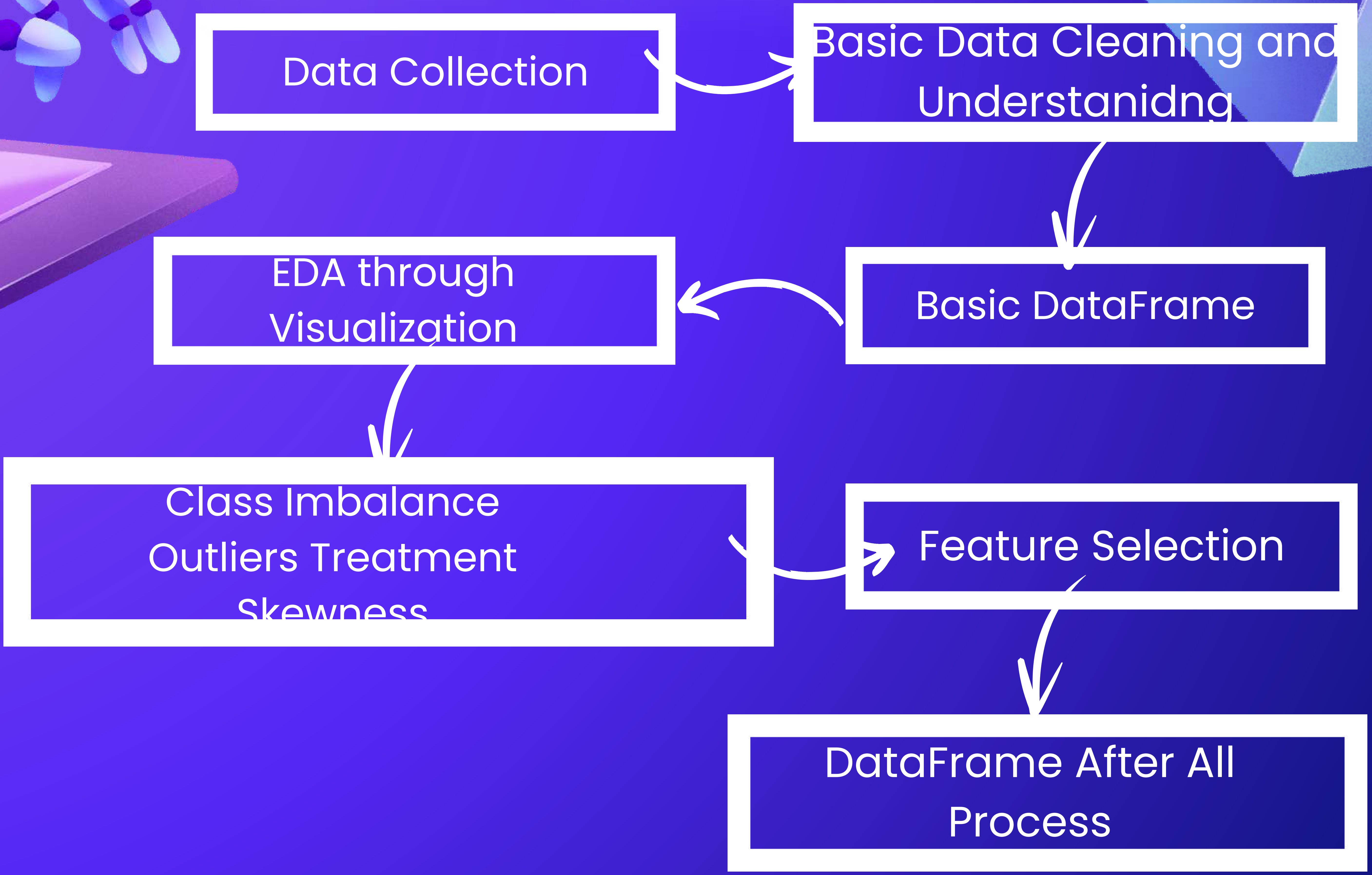
INTRODUCTION



Employee turnover, or the rate at which employees leave a company and are replaced by new hires, is a critical metric for organizations across industries.

In this model we will be deplloying supervised Machine learning techniques to prefict the turnover

FLOW OF THE PROJECT



Data Collection

Basic Data Cleaning and Understanidng

Basic DataFrame

EDA through Visualization

Class Imbalance
Outliers Treatment
Skewness

Feature Selection

DataFrame After All Process

ABOUT DATASET

DataSet: Real dataset shared from Edward Abushkin’s blog used to predict an Employee's risk of quitting.
Columns: 10

Features	**"department" , "promoted" ,"review" ,"projects" ,”salary","tenure","satisfaction" ,"bonus" ,"avg_hrs_month" , "left"**
Target	“Left”
No.of Rows	9540

Problem Statement

Predict if employee will leave company.

BASIC CLEANING:

Null Values	0
DATA TYPE OBJECT	"department”,”salary”,”left”,“Promoted”,”bonus
DATA TYPE NUMERIC	“review“,”projects”,”tenure”,satisfaction”,”avg_hrs_month”.

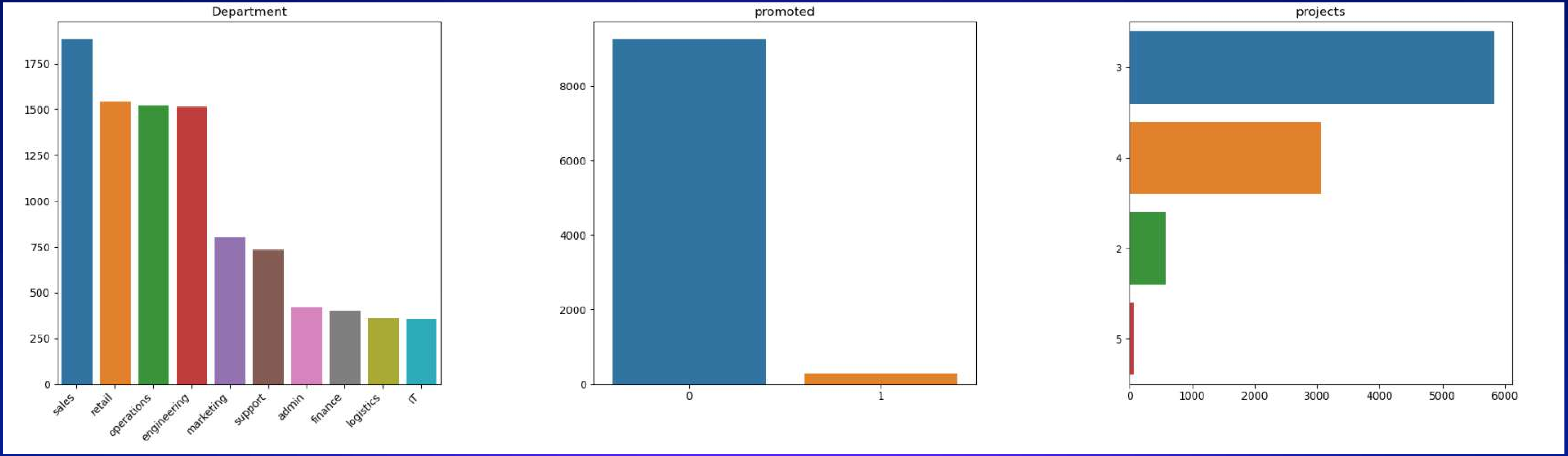
DATA TYPE CONVERSION USING LABEL ENCODER

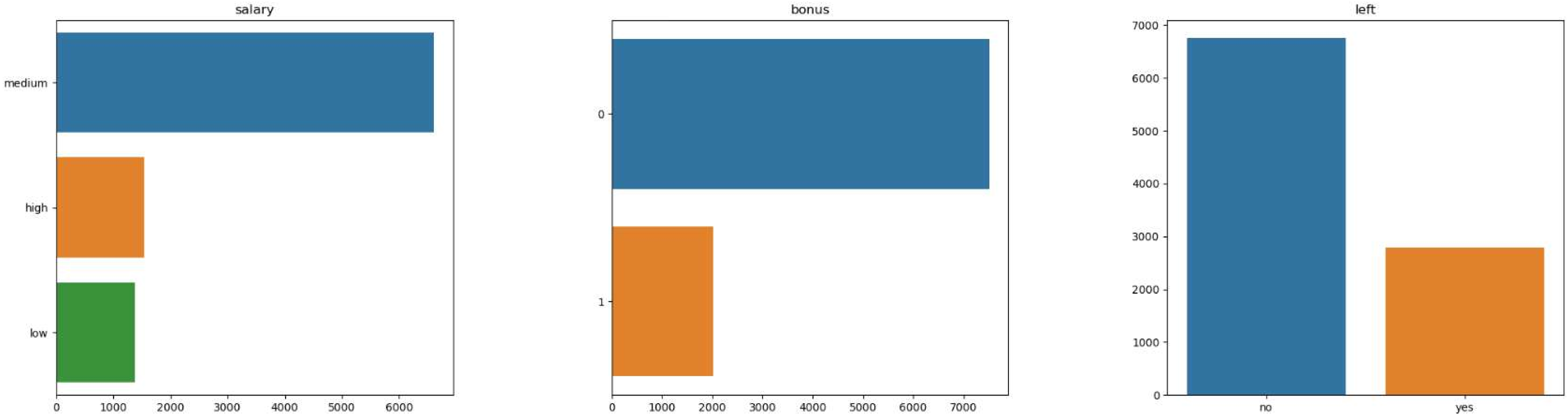
01

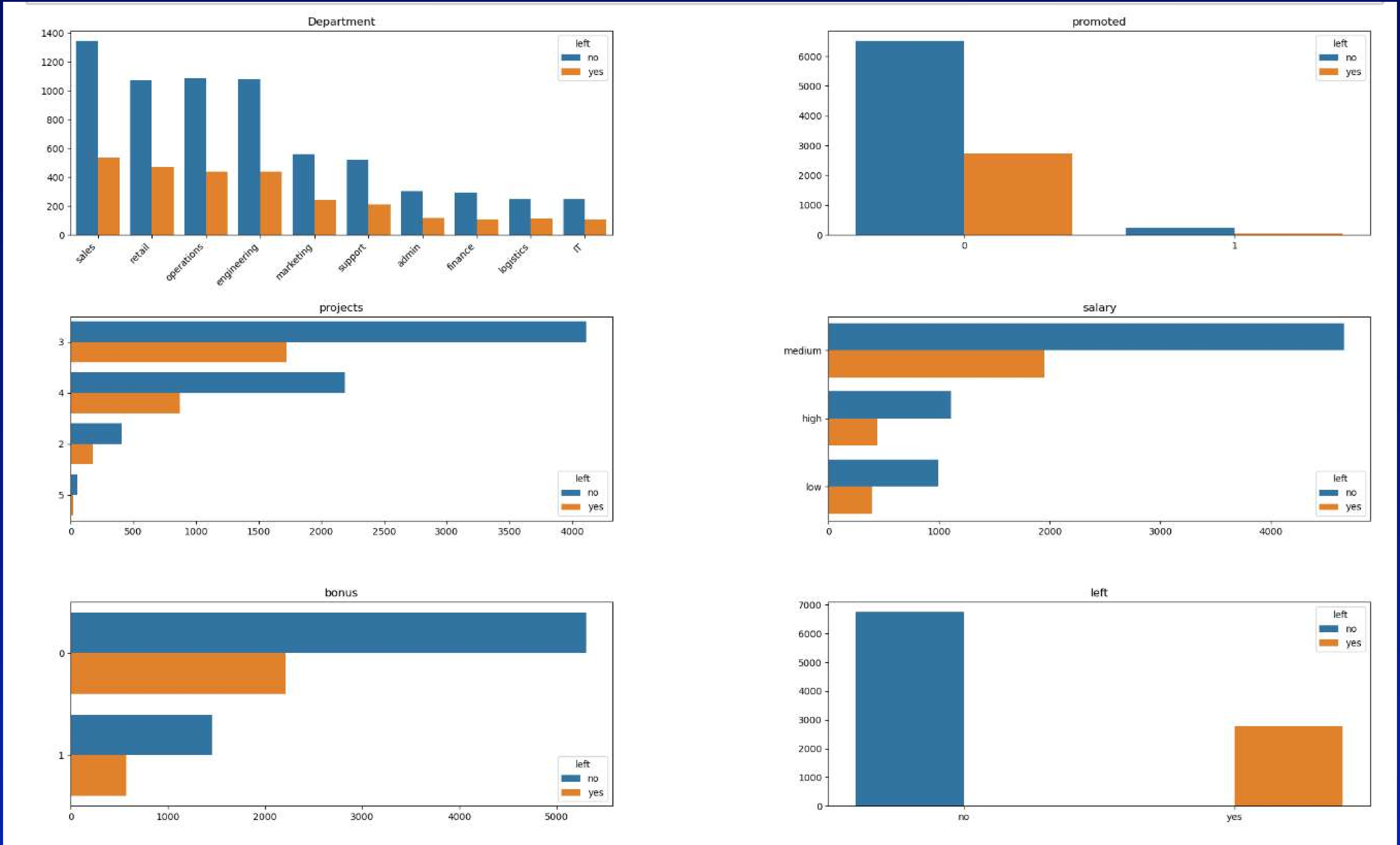
BASIC DATAFRAME

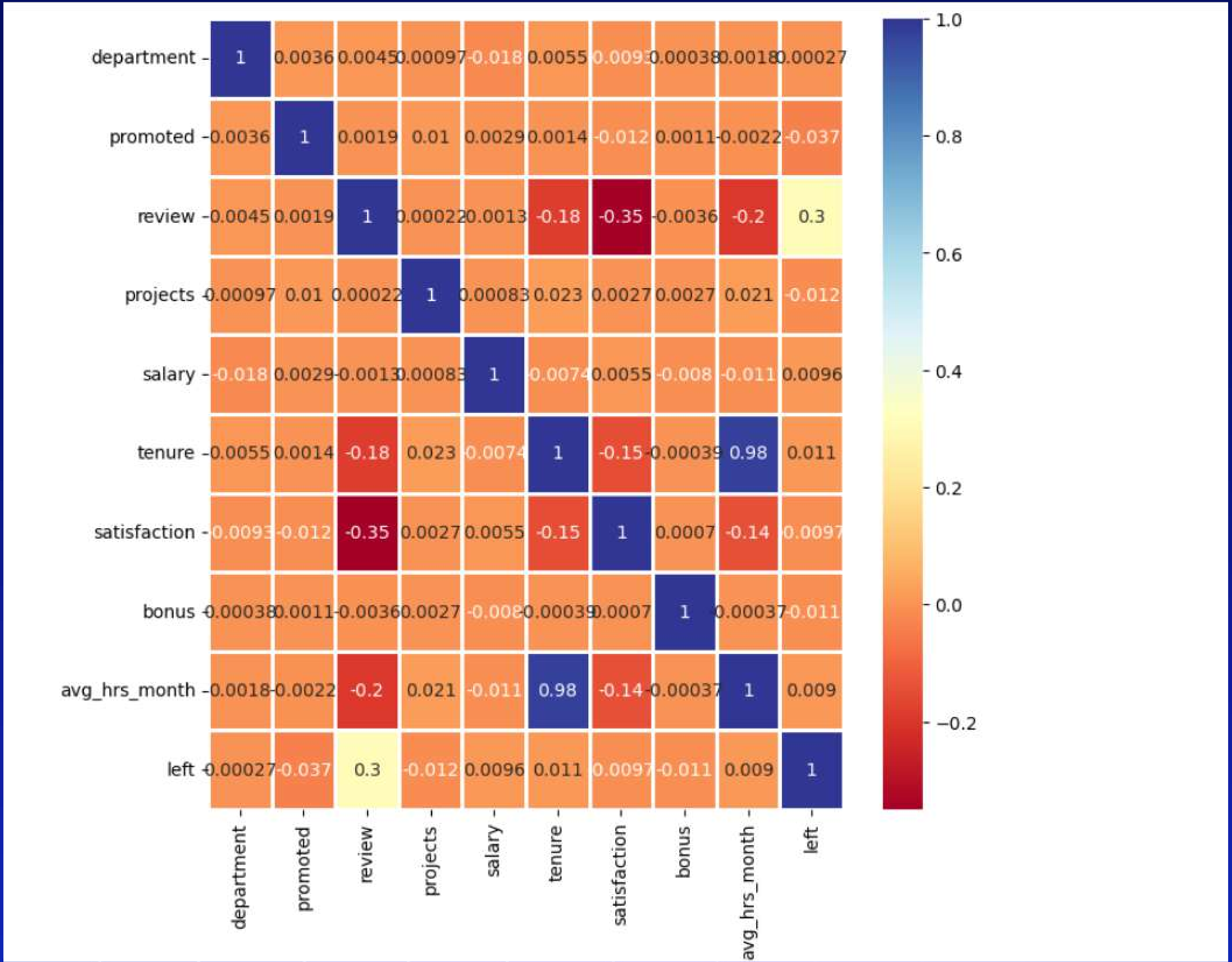
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	72.327044	0.726810	0.966165	0.264624
Decision Tree	81.813417	0.860844	0.874436	0.694811
Random Forest	85.691824	0.872702	0.927820	0.747456
KNN	73.322851	0.772757	0.874436	0.481142
SVM	69.706499	0.697065	1.000000	0.000000
XGBoost	85.534591	0.879683	0.918045	0.748634

02 VISUALIZATION




salary
medium
high
low
0
1000
2000
3000
4000
5000
6000
bonus
0
1
0
1000
2000
3000
4000
5000
6000
7000
left
7000
6000
5000
4000
3000
2000
1000
0
no
yes


Department
1400
1200
1000
800
600
400
200
0
sales
retail
operations
engineering
marketing
support
admin
finance
logistics
IT
left
no
yes
promoted
6000
5000
4000
3000
2000
1000
0
0
1
left
no
yes
projects
3
4
2
5
0
500
1000
1500
2000
2500
3000
3500
4000
left
no
yes
salary
medium
high
low
0
1000
2000
3000
4000
left
no
yes
bonus
0
1
0
1000
2000
3000
4000
5000
left
no
yes
left
7000
6000
5000
4000
3000
2000
1000
0
no
yes
left
no
yes


department 1 0.0036 0.0045 0.00097 -0.018 0.0055 -0.0093 0.00038 0.0018 0.00027
promoted 0.0036 1 0.0019 0.01 0.0029 0.0014 -0.012 0.0011 -0.0022 -0.037
review 0.0045 0.0019 1 0.00022 0.0013 -0.18 -0.35 -0.0036 -0.2 0.3
projects 0.00097 0.01 0.00022 1 0.00083 0.023 0.0027 0.0027 0.021 -0.012
salary -0.018 0.0029 -0.0013 0.00083 1 -0.0074 0.0055 -0.008 -0.011 0.0096
tenure 0.0055 0.0014 -0.18 0.023 -0.0074 1 -0.15 -0.00039 0.98 0.011
satisfaction -0.0093 -0.012 -0.35 0.0027 0.0055 -0.15 1 0.0007 -0.14 -0.0097
bonus 0.00038 0.0011 -0.0036 0.0027 -0.008 -0.00039 0.0007 1 0.00037 -0.011
avg_hrs_month 0.0018 -0.0022 -0.2 0.021 -0.011 0.98 -0.14 -0.00037 1 0.009
left 0.00027 -0.037 0.3 -0.012 0.0096 0.011 -0.0097 -0.011 0.009 1
department
promoted
review
projects
salary
tenure
satisfaction
bonus
avg_hrs_month
left
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

03

CLASS IMBALANCE

TARGET COLUMN “LEFT “

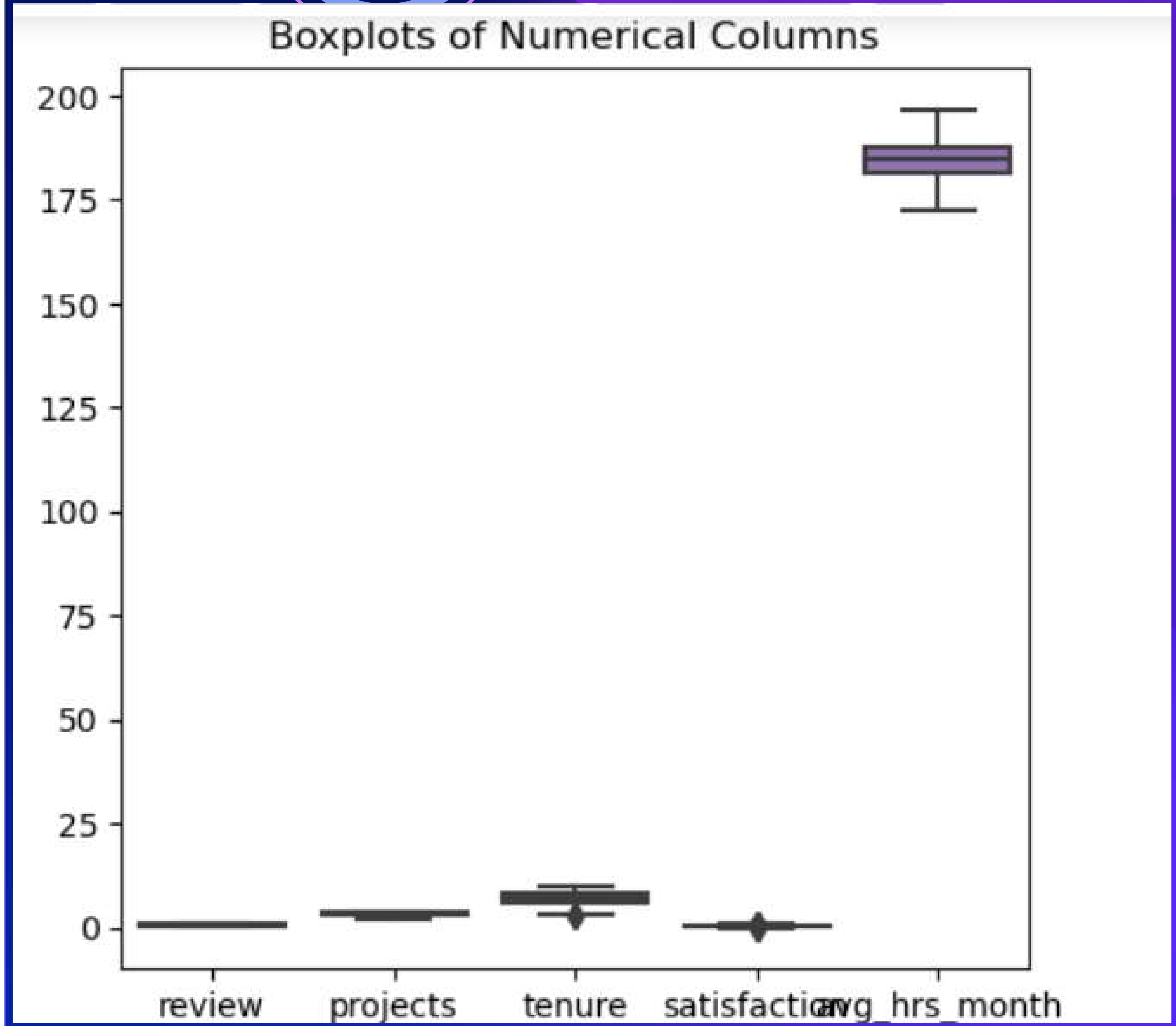
HAD CLASS IMBALANCE

ACTION TAKEN : APPLIED SMOTE TECHNIQUE

"LEFT" - "YES" IF THE EMPLOYEE ENDED UP LEAVING, "NO" OTHERWISE.

Instance	No.of “NO”	**No.of”YES”**
Before	6756	2784
After	6756	6756

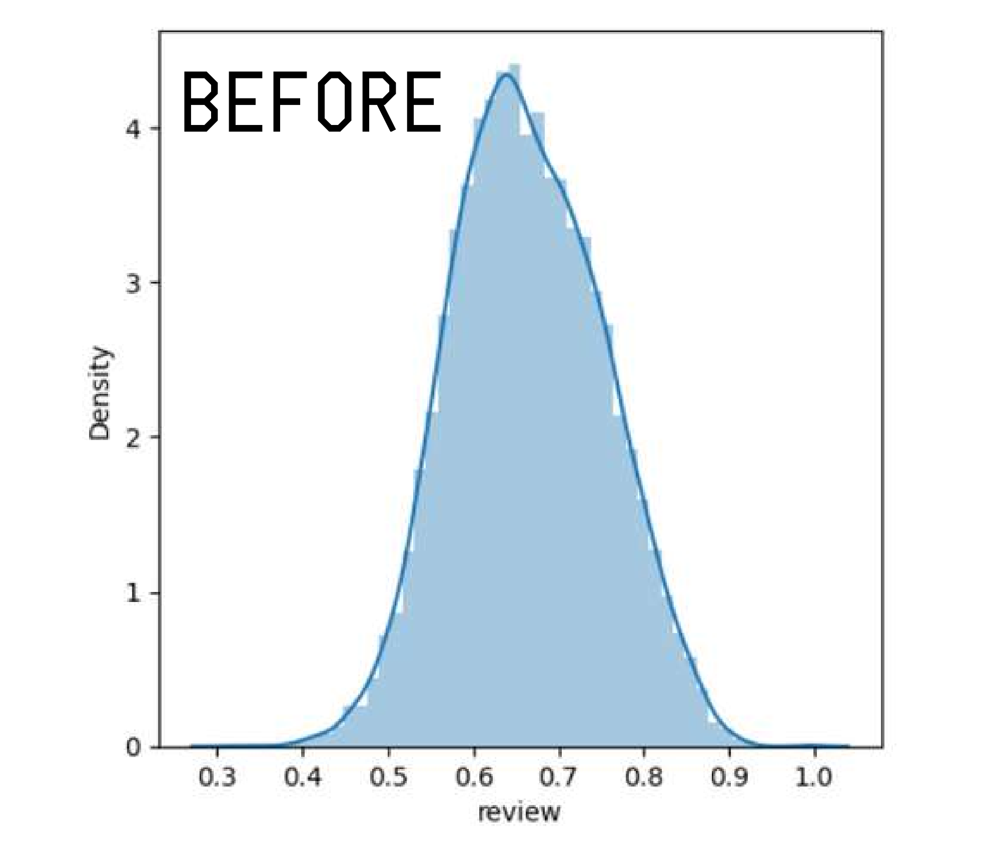
04 OUTLIERS

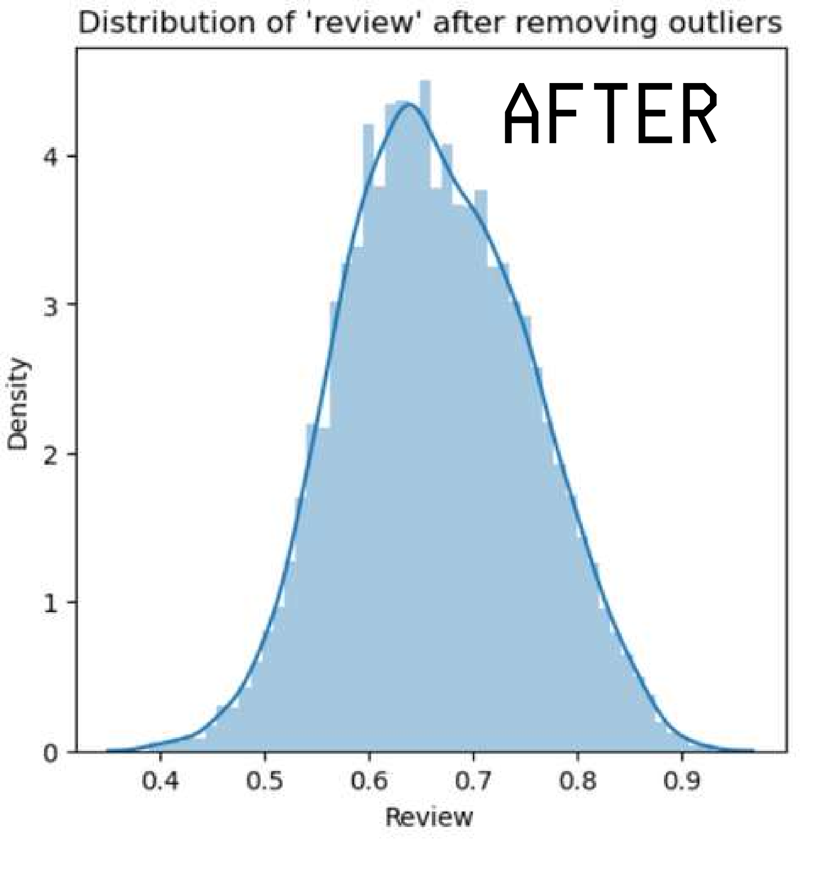


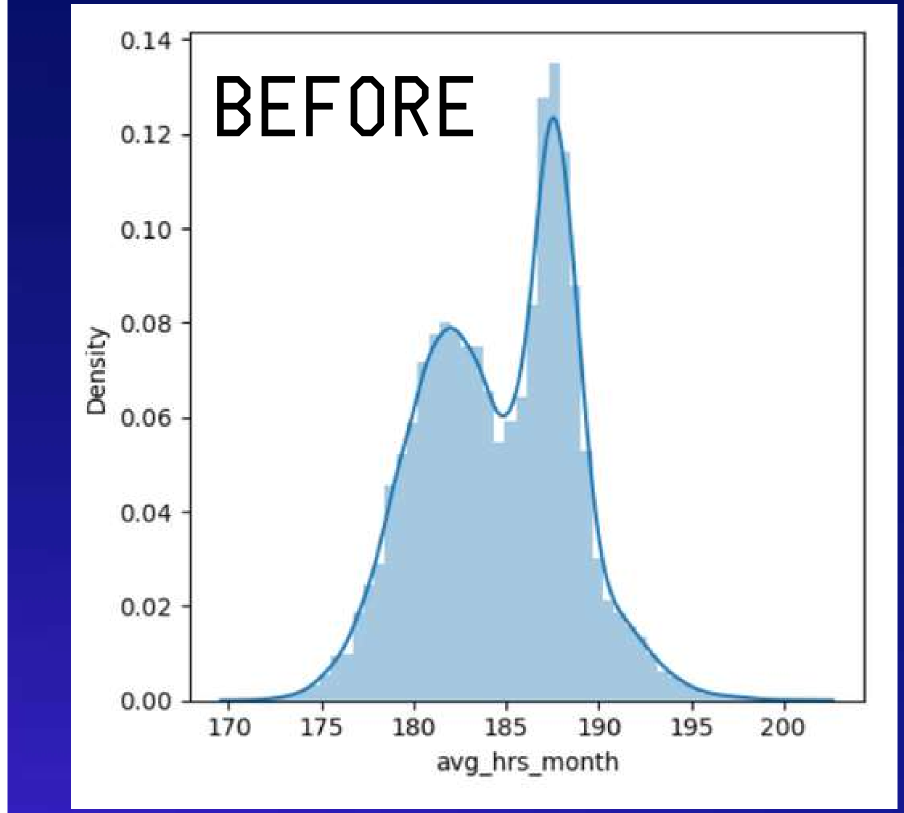
NUMERICAL COLUMNS WHICH REQUIRE OUTLIERS TREATMENT :

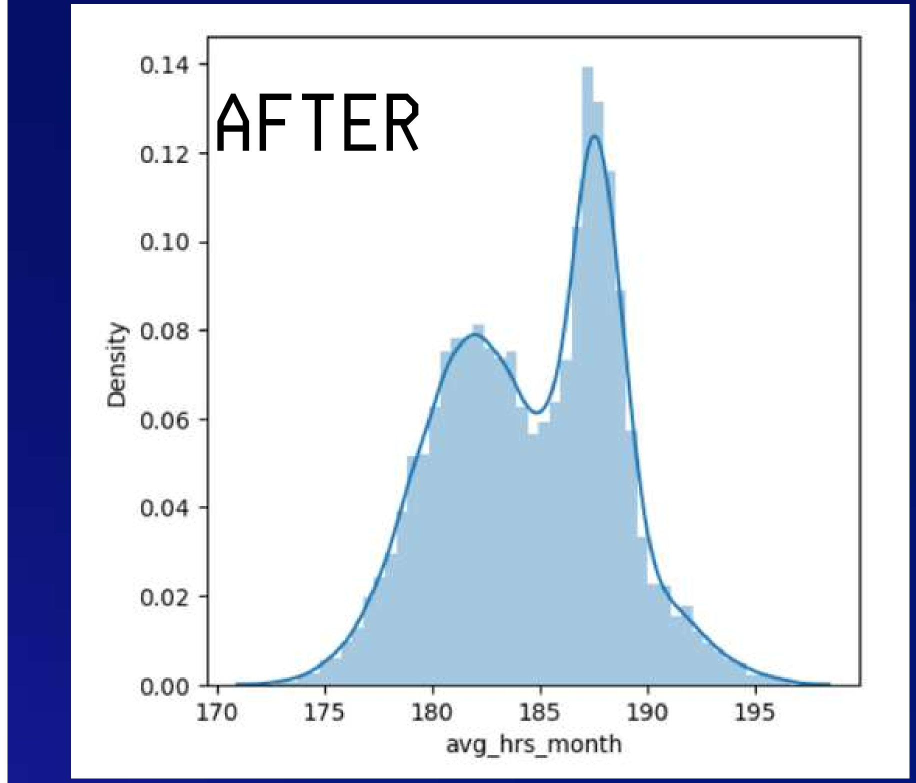
'REVIEW','PROJECTS','TENURE",'AVG_HRS_MONTH'

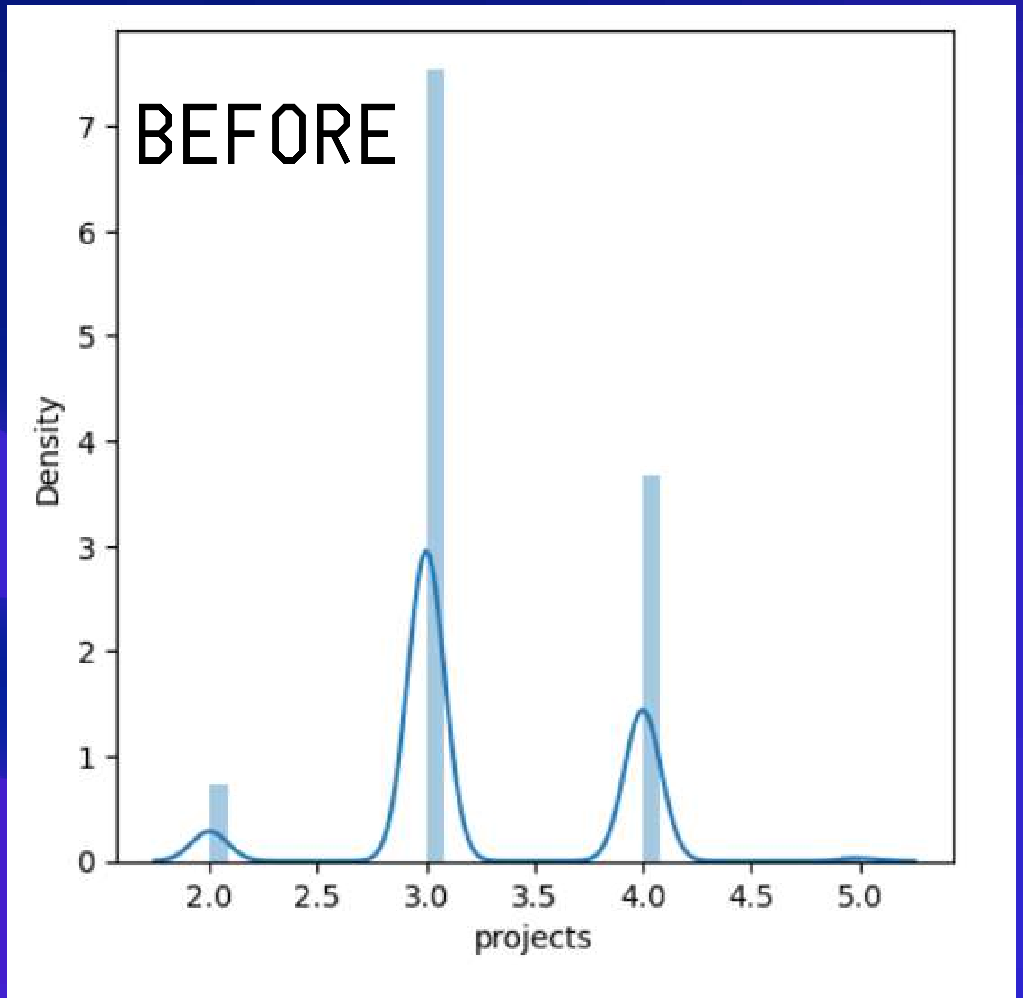
Type of Distribution	Method Used	Columns
Normal	Standard Deviation	“review”,”project”,”tenure”,”avg_hrs_month”
Skewed	IQR	NA

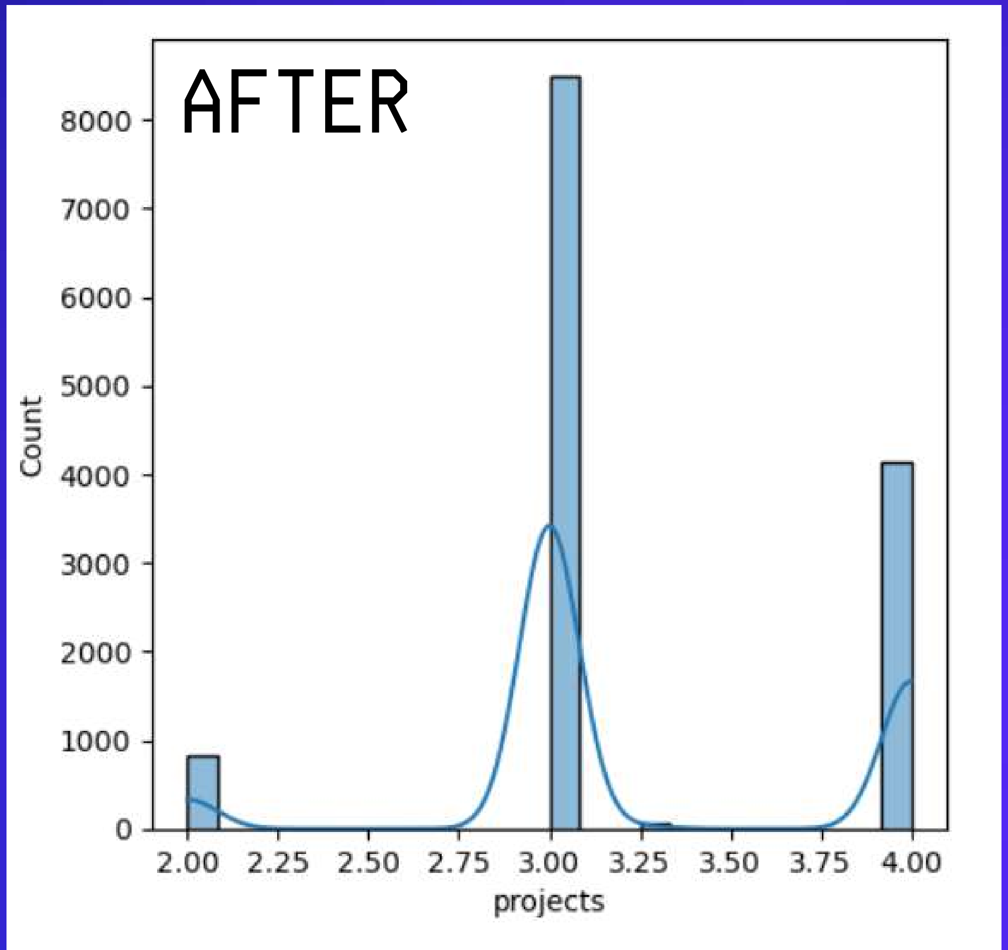

BEFORE
Density
0
1
2
3
4
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
review

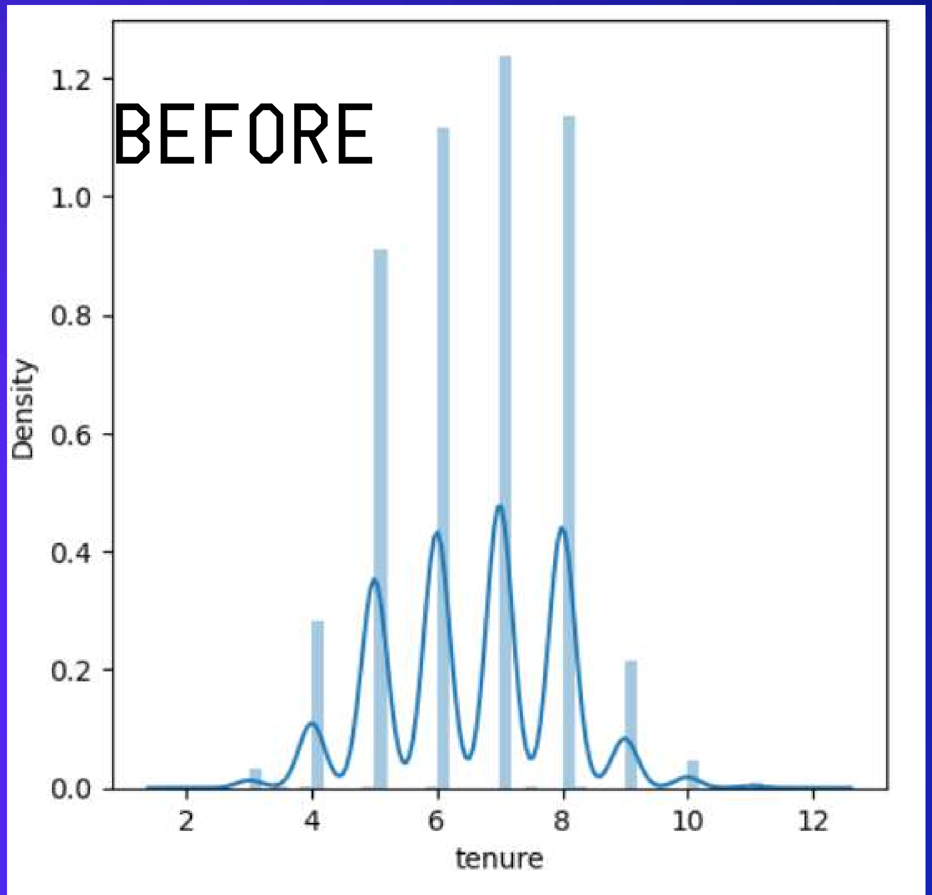

Distribution of 'review' after removing outliers
AFTER
Density
0
1
2
3
4
0.4
0.5
0.6
0.7
0.8
0.9
Review

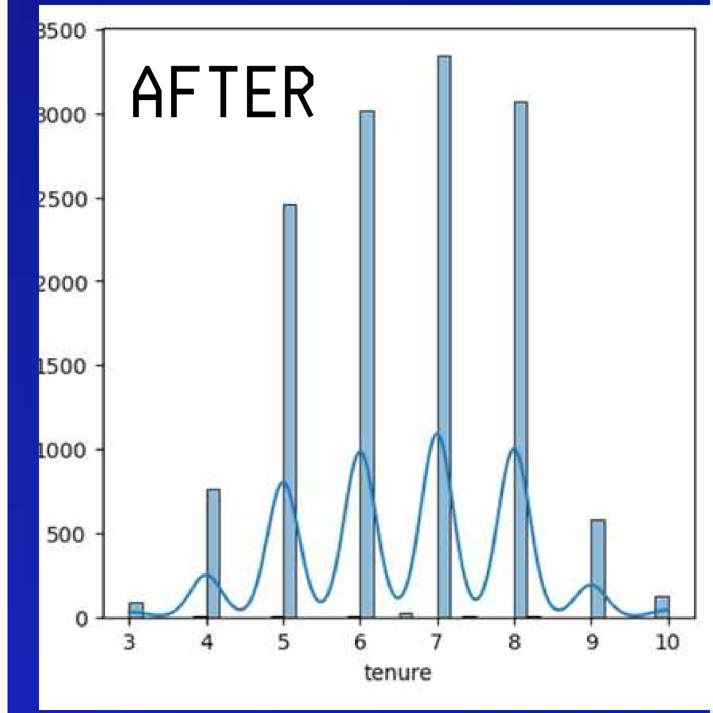

BEFORE
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
170
175
180
185
190
195
200
avg_hrs_month


AFTER
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
170
175
180
185
190
195
avg_hrs_month


BEFORE
Density
0
1
2
3
4
5
6
7
2.0
2.5
3.0
3.5
4.0
4.5
5.0
projects


AFTER
Count
0
1000
2000
3000
4000
5000
6000
7000
8000
2.00
2.25
2.50
2.75
3.00
3.25
3.50
3.75
4.00
projects


BEFORE
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
2
4
6
8
10
12
tenure


AFTER
3500
3000
2500
2000
1500
1000
500
0
3
4
5
6
7
8
9
10
tenure

05

SKEWNESS

Column	Skewness	Skewness Treatment Required
department	-0.526152	NO
promoted	6.533033	NO
review	0.153407	NO
projects	0.023403	NO
salary	-1.237002	NO
tenure	-0.099847	NO
satisfaction	0.099041	NO
bonus	1.554054	NO
avg_hrs_month	-0.119797	NO

06

DATAFRAME_2

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	53.015168	0.691834	0.664693	0.558414
Decision Tree	80.096189	0.842502	0.827535	0.802496
Random Forest	85.423603	0.884384	0.871947	0.85482
KNN	74.953755	0.797524	0.667654	0.764522
SVM	49.981502	0.498144	0.993338	0.000000
XGBoost	83.3518	0.862324	0.874625	0.834680
AdaBoost	69.367370	0.811979	0.792746	0.675549

DATAFRAME_1 V/S DATAFRAME_2

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	72.327044	0.726810	0.966165	0.264624
Decision Tree	81.813417	0.860844	0.874436	0.694811
Random Forest	85.691824	0.872702	0.927820	0.747456
KNN	73.322851	0.772757	0.874436	0.481142
SVM	69.706499	0.697065	1.000000	0.000000
XGBoost	85.534591	0.879683	0.918045	0.748634

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	53.015168	0.691834	0.664693	0.558414
Decision Tree	80.096189	0.842502	0.827535	0.802496
Random Forest	85.423603	0.884384	0.871947	0.85482
KNN	74.953755	0.797524	0.667654	0.764522
SVM	49.981502	0.498144	0.993338	0.000000
XGBoost	83.3518	0.862324	0.874625	0.834680
AdaBoost	69.367370	0.811979	0.792746	0.675549

FEATURE SELECTION

We see promoted Salary projects and Bonus are least important

Importance	Column
0.049497	department
0.003577	promoted
0.260457	review
0.021433	projects
0.018584	salary
0.065972	tenure
0.301302	satisfaction
0.012011	bonus
0.267167	avg_hrs_month

DATA FRAME AFTER FEATURE SELECTION

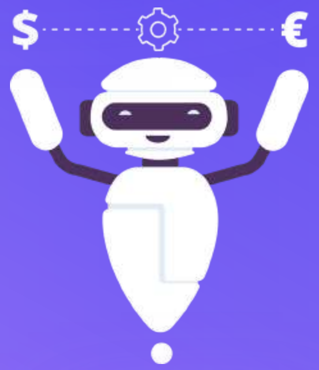
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	53.570107	0.530918	0.629164	0.487127
Decision Tree	78.727340	0.794852	0.777202	0.789454
Random Forest	83.980762	0.848187	0.831236	0.841450
KNN	80.133185	0.821853	0.768320	0.808556
SVM	46.281909	0.495896	0.983716	0.004115
XGBoost	83.166852	0.829756	0.834077	0.833150
AdaBoost	68.183500	0.693796	0.769800	0.661684

Model	BEFORE Accuracy	Precision	Recall	F1-Score	AFTER Accuracy	Precision	Recall	F1-Score
Logistic Regression	53.015168	0.691834	0.664693	0.558414	53.570107	0.530918	0.629164	0.487127
Decision Tree	80.096189	0.842502	0.827535	0.802496	78.727340	0.794852	0.777202	0.789454
Random Forest	85.423603	0.884384	0.871947	0.85482	83.980762	0.848187	0.831236	0.841450
KNN	74.953755	0.797524	0.667654	0.764522	80.133185	0.821853	0.768320	0.808556
SVM	49.981502	0.498144	0.993338	0.000000	46.281909	0.495896	0.983716	0.004115
XGBoost	83.3518	0.862324	0.874625	0.834680	83.166852	0.829756	0.834077	0.833150
AdaBoost	69.367370	0.811979	0.792746	0.675549	68.183500	0.693796	0.769800	0.661684

PROJECT

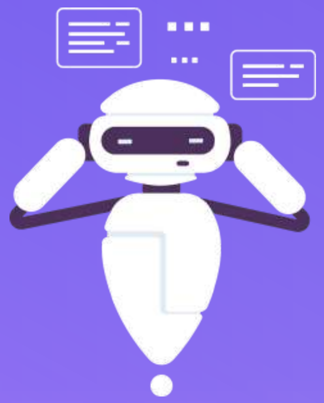
DEMERITS

- Error Chances
- Dependence on Data Quality



FACTS

The Accurecy which ever atmost we get we cannot tell it 100% that employee will leave or stay as this thing have much more dimensions in reality



BASED ON DATASET

Conclusion :

Random Forest is model which is giving the best accurecy outof Linear, Decison tree , Random forest , SVM,KNN,,XGBoost and Ada boost

THANK YOU!