

Analysis of Employee Retention rate

Internship in eSparsh Technologies Pvt Ltd(Bangalore)

Table of Contents

- Introduction
- Abstract
- The dataset
- Data Preprocessing
- Training the Dataset
- Choosing the Machine Learning Model
- About the Machine Learning Model
- Deploy the Model
- Visualization
- Conclusion

Abstract

Employee retention is a critical concern for organizations, and predicting retention rates can help proactively identify factors influencing attrition. In this project, we propose a machine learning-based approach to predict employee retention rates using historical employee data. Through data preprocessing, feature engineering, and training various machine learning models, we evaluate and optimize their performance. The selected model is deployed in a production environment for real-time predictions. This project aims to provide valuable insights into factors influencing retention, enabling organizations to implement targeted strategies for improving employee satisfaction and reducing attrition, ultimately leading to enhanced talent management and organizational performance.

Introduction

- Employee retention is a critical concern for organizations as high turnover rates can incur significant costs and disrupt business operations.
 - To address this challenge, organizations are increasingly adopting data-driven approaches to gain insights into factors influencing employee retention.
 - This project aims to develop a machine learning-based model for predicting employee retention rates using historical employee data.
 - By analysing factors such as job satisfaction, age, promotion history, work-life balance, and employee demographics, organizations can understand the drivers of attrition and implement targeted strategies for improving retention.
 - The project involves stages such as data collection from the HR database, data preprocessing to handle missing values and outliers, and feature engineering to extract meaningful insights.
 - Machine learning algorithms including logistic regression, decision trees, random forests, and XGB Classifier are trained and evaluated to identify the best model for predicting retention rates.
- enhance employee satisfaction, engagement, and overall talent management.
- By leveraging machine learning, organizations can gain a deeper understanding of retention factors, leading to improved organizational performance and a positive work environment.

the 1990s, the number of people in the United States who are 65 years of age or older has increased by 50 percent, and the number of people 75 years of age or older has increased by 100 percent. The number of people 85 years of age or older has increased by 200 percent. The number of people 95 years of age or older has increased by 400 percent. The number of people 100 years of age or older has increased by 1,000 percent. The number of people 105 years of age or older has increased by 2,000 percent. The number of people 110 years of age or older has increased by 4,000 percent. The number of people 115 years of age or older has increased by 8,000 percent. The number of people 120 years of age or older has increased by 16,000 percent. The number of people 125 years of age or older has increased by 32,000 percent. The number of people 130 years of age or older has increased by 64,000 percent. The number of people 135 years of age or older has increased by 128,000 percent. The number of people 140 years of age or older has increased by 256,000 percent. The number of people 145 years of age or older has increased by 512,000 percent. The number of people 150 years of age or older has increased by 1,024,000 percent. The number of people 155 years of age or older has increased by 2,048,000 percent. The number of people 160 years of age or older has increased by 4,096,000 percent. The number of people 165 years of age or older has increased by 8,192,000 percent. The number of people 170 years of age or older has increased by 16,384,000 percent. The number of people 175 years of age or older has increased by 32,768,000 percent. The number of people 180 years of age or older has increased by 65,536,000 percent. The number of people 185 years of age or older has increased by 131,072,000 percent. The number of people 190 years of age or older has increased by 262,144,000 percent. The number of people 195 years of age or older has increased by 524,288,000 percent. The number of people 200 years of age or older has increased by 1,048,576,000 percent. The number of people 205 years of age or older has increased by 2,097,152,000 percent. The number of people 210 years of age or older has increased by 4,194,304,000 percent. The number of people 215 years of age or older has increased by 8,388,608,000 percent. The number of people 220 years of age or older has increased by 16,777,216,000 percent. The number of people 225 years of age or older has increased by 33,554,432,000 percent. The number of people 230 years of age or older has increased by 67,108,864,000 percent. The number of people 235 years of age or older has increased by 134,217,728,000 percent. The number of people 240 years of age or older has increased by 268,435,456,000 percent. The number of people 245 years of age or older has increased by 536,870,912,000 percent. The number of people 250 years of age or older has increased by 1,073,741,824,000 percent. The number of people 255 years of age or older has increased by 2,147,483,648,000 percent. The number of people 260 years of age or older has increased by 4,294,967,296,000 percent. The number of people 265 years of age or older has increased by 8,589,934,592,000 percent. The number of people 270 years of age or older has increased by 17,179,869,184,000 percent. The number of people 275 years of age or older has increased by 34,359,738,368,000 percent. The number of people 280 years of age or older has increased by 68,719,476,736,000 percent. The number of people 285 years of age or older has increased by 137,438,953,472,000 percent. The number of people 290 years of age or older has increased by 274,877,906,944,000 percent. The number of people 295 years of age or older has increased by 549,755,813,888,000 percent. The number of people 300 years of age or older has increased by 1,099,511,627,776,000 percent. The number of people 305 years of age or older has increased by 2,199,023,255,552,000 percent. The number of people 310 years of age or older has increased by 4,398,046,511,104,000 percent. The number of people 315 years of age or older has increased by 8,796,093,022,208,000 percent. The number of people 320 years of age or older has increased by 17,592,186,044,416,000 percent. The number of people 325 years of age or older has increased by 35,184,372,088,832,000 percent. The number of people 330 years of age or older has increased by 70,368,744,177,664,000 percent. The number of people 335 years of age or older has increased by 140,737,488,355,328,000 percent. The number of people 340 years of age or older has increased by 281,474,976,710,656,000 percent. The number of people 345 years of age or older has increased by 562,949,953,421,312,000 percent. The number of people 350 years of age or older has increased by 1,125,899,906,842,624,000 percent. The number of people 355 years of age or older has increased by 2,251,799,813,685,248,000 percent. The number of people 360 years of age or older has increased by 4,503,599,627,370,496,000 percent. The number of people 365 years of age or older has increased by 9,007,199,254,740,992,000 percent. The number of people 370 years of age or older has increased by 18,014,398,509,481,984,000 percent. The number of people 375 years of age or older has increased by 36,028,797,018,963,968,000 percent. The number of people 380 years of age or older has increased by 72,057,594,037,927,936,000 percent. The number of people 385 years of age or older has increased by 144,115,188,075,855,872,000 percent. The number of people 390 years of age or older has increased by 288,230,376,151,711,744,000 percent. The number of people 395 years of age or older has increased by 576,460,752,303,423,488,000 percent. The number of people 400 years of age or older has increased by 1,152,921,504,606,846,976,000 percent. The number of people 405 years of age or older has increased by 2,305,843,009,213,693,952,000 percent. The number of people 410 years of age or older has increased by 4,611,686,018,427,387,904,000 percent. The number of people 415 years of age or older has increased by 9,223,372,036,854,775,808,000 percent. The number of people 420 years of age or older has increased by 18,446,744,073,709,551,616,000 percent. The number of people 425 years of age or older has increased by 36,893,488,147,419,103,232,000 percent. The number of people 430 years of age or older has increased by 73,786,976,294,838,206,464,000 percent. The number of people 435 years of age or older has increased by 147,573,952,589,676,412,928,000 percent. The number of people 440 years of age or older has increased by 295,147,905,179,352,825,856,000 percent. The number of people 445 years of age or older has increased by 590,295,810,358,705,651,712,000 percent. The number of people 450 years of age or older has increased by 1,180,591,620,717,411,303,424,000 percent. The number of people 455 years of age or older has increased by 2,361,183,241,434,822,606,848,000 percent. The number of people 460 years of age or older has increased by 4,722,366,482,869,645,213,696,000 percent. The number of people 465 years of age or older has increased by 9,444,732,965,739,290,427,392,000 percent. The number of people 470 years of age or older has increased by 18,889,465,931,478,580,854,784,000 percent. The number of people 475 years of age or older has increased by 37,778,931,862,957,161,709,568,000 percent. The number of people 480 years of age or older has increased by 75,557,863,725,914,323,419,136,000 percent. The number of people 485 years of age or older has increased by 151,115,727,451,828,646,838,272,000 percent. The number of people 490 years of age or older has increased by 302,231,454,903,657,293,676,544,000 percent. The number of people 495 years of age or older has increased by 604,462,909,807,314,587,353,088,000 percent. The number of people 500 years of age or older has increased by 1,208,925,819,614,629,174,706,176,000 percent. The number of people 505 years of age or older has increased by 2,417,851,639,229,258,349,412,352,000 percent. The number of people 510 years of age or older has increased by 4,835,703,278,458,516,698,824,704,000 percent. The number of people 515 years of age or older has increased by 9,671,406,556,917,033,397,649,408,000 percent. The number of people 520 years of age or older has increased by 19,342,813,113,834,066,795,298,816,000 percent. The number of people 525 years of age or older has increased by 38,685,626,227,668,133,590,597,632,000 percent. The number of people 530 years of age or older has increased by 77,371,252,455,336,267,181,195,264,000 percent. The number of people 535 years of age or older has increased by 154,742,504,910,672,534,362,390,528,000 percent. The number of people 540 years of age or older has increased by 309,485,009,821,345,068,724,781,056,000 percent. The number of people 545 years of age or older has increased by 618,970,019,642,690,137,449,562,112,000 percent. The number of people 550 years of age or older has increased by 1,237,940,039,285,380,274,899,124,224,000 percent. The number of people 555 years of age or older has increased by 2,475,880,078,570,760,549,798,248,448,000 percent. The number of people 560 years of age or older has increased by 4,951,760,157,141,521,099,596,496,896,000 percent. The number of people 565 years of age or older has increased by 9,903,520,314,283,042,199,193,993,792,000 percent. The number of people 570 years of age or older has increased by 19,807,040,628,566,084,398,387,

EMPCODE	Name	BirthDate	Age	Gender	Department	Designation	JoiningDate	LeavingDate	WORKPERIOD IN MONTHS	WORKPERIOD IN DAYS	Work Period	Retention	Reason	PermanentState
0259	Manikanta P C	07-Mar-82	41.345632	Male	Engineering	Associate Engineer	14-Nov-05	13-Jun-19	165	4959.00	13 years-6 months	NO	Higher Education	
0809	Karunakara Reddy M	09-May-82	41.173142	Male	Engineering	Lead Engineer	14-Feb-11	12-Oct-18	93	2797.00	7 years-7 months	NO	Career Growth	Karnataka
1043	NAVYA THADATIIL	10-Oct-83	39.751519	Female	HR & Admin	Deputy Manager	28-Nov-11	24-Apr-18	78	2339.00	6 years-4 months	NO	Career Growth	Karnataka
1044	SUBBAREDDY K	23-Nov-77	45.631015	Male	Engineering	Associate Engineer	05-Dec-11	14-Oct-21	120	3601.00	9 years-10 months	NO	Personal Reason	Karnataka
1049	LIJIN SHAJI	31-Jan-82	41.44146	Male	Engineering	Associate Senior Manager	09-Jan-12	21-May-21	114	3420.00	9 years-4 months	NO	Personal Reason	Kerala
1063	YOGESHAY S	19-Oct-85	37.726832	Male	Engineering	Associate Senior Manager	27-Feb-12	06-Feb-23	133	3997.00	10 years-11 months	NO	Better Opportunity	Karnataka
1086	BASAVANN EVVA CHOUDHARI	22-Jun-86	37.05267	Female	Engineering	Lead Engineer	11-Jun-12	01-Jun-20	97	2912.00	7 years-11 months	NO	Personal Reason	
1090	GURURAJAN R	31-Dec-85	37.526959	Male	Engineering	Associate Engineer	02-Jul-12	10-Jan-19	79	2383.00	6 years-6 months	NO	Better Opportunity	
1100	PRADEEP KUMAR B	01-Jan-85	38.523589	Male	Engineering	Lead Engineer	16-Jul-12	16-Jul-18	73	2191.00	6 years-0 months	NO	Career Growth	Karnataka
1118	VIKAS G HEGDE	07-Oct-82	40.759713	Male	Engineering	Senior Manager	17-Sep-12			-41169.00		YES		Karnataka
1172	MADHUSUDAN SAHOO	20-Jun-87	36.058162	Male	Engineering	Associate Architect	11-Mar-13	30-Jul-21	102	3063.00	8 years-4 months	NO	Career Growth	Odisha
1180	VIJAYALAXMI S ULVEKAR	10-Nov-85	37.666596	Female	Engineering	Architect	01-Apr-13			-41365.00		YES		
1182	VANGARA NAGAMANI	10-Jun-81	42.085514	Female	Engineering	Architect	01-Apr-13	06-Dec-21	106	3171.00	8 years-8 months	NO	Better Opportunity	Andhra Pradesh
1209	N HARIKRISHNARAJAN	09-Oct-74	48.754244	Male	Engineering	Associate Engineer	01-Jul-13	14-Jun-19	72	2174.00	5 years-11 months	NO	Career Growth	Tamil Nadu
1217	ANANTH K L	07-Jun-88	35.091034	Male	Engineering	Associate Engineer	22-Jul-13	22-Nov-18	65	1949.00	5 years-4 months	NO	Career Growth	Karnataka
1228	ASHWINI K.G	09-Mar-91	32.340164	Female	Engineering	Senior Software Engineer	19-Aug-13	31-Dec-18	65	1960.00	5 years-4 months	NO	Career Growth	
1230	PAVAN KUNTE A	13-Mar-91	32.329213	Male	Engineering	Senior Software Engineer	19-Aug-13	23-Nov-18	64	1922.00	5 years-3 months	NO	Career Growth	Karnataka
1233	KAVYA R	07-May-92	31.175907	Female	Engineering	Senior Software Engineer	19-Aug-13	14-Aug-18	61	1821.00	4 years-11 months	NO	Career Growth	Karnataka
1234	DHANARAJ	01-Jul-85	37.751519	Male	Engineering	Associate Engineer	14-Nov-05	13-Jun-19	165	4959.00	13 years-6 months	NO	Higher Education	

Attribute Description

Attribute	Description	Type
Age	Age of Employees	Numerical Discrete
Gender	Gender of the person	Categorical
Department	1-Engineer,2-IT,3-Finance,.....	Categorical
Designation	1- Associate Engineer,2-Architect,3-Assistant manger,.....	Categorical
Work period in Months	Numeric	Discrete
Joining Date	Date	dd//mm//yyyy
Leaving Date	Date	dd//mm//yyyy
Retention	1- Yes, 0 - No	Categorical

Data Preprocessing

- Removing the unwanted columns/fields in the dataset.
- Checking for Null values.
- Fill the Null values.
 - Data Type - Numerical - Use Mean / Median
 - Data Type - Text - Mode
- Change the categorical values to numerical.
 - Label encoder
 - Word to vector

Training and testing the Data into Model

```
[ ] from sklearn.model_selection import train_test_split
    from sklearn.ensemble import GradientBoostingClassifier
    from sklearn.metrics import accuracy_score
    from sklearn.preprocessing import LabelEncoder
```

```
[ ] # Splitting the dataset into input features (X) and target variable (y)
    X = data[['Age', 'Workperiod in Months', 'Gender', 'Department', 'Designation']]
    y = data['Retention']
```

```
[ ] X_encoded = pd.get_dummies(X)
```

```
▶ # Convert the target variable to numerical labels
    label_encoder = LabelEncoder()
    y = label_encoder.fit_transform(y)
    print(y)
```

```
↳ [0 0 0 ... 1 0 1]
```

```
[ ] print(X_encoded)
```

```
[ ] # Splitting the data into training and testing sets
    X_train, X_test, y_train, y_test = train_test_split(X_encoded, y, test_size=0.2, random_state=42), stratify = y
```


Choose the Machine Learning Model

For the above features and the target variable we can use the following model,

- GradientBoostingClassifier - 0.794
- LogisticRegression - 0.686
- XGBClassifier - 0.8

Hence here we have high accuracy in **XGB Classifier**.

GradientBoostingClassifier

```
▶ # Create an instance of GradientBoostingClassifier
model = GradientBoostingClassifier()

# Train the model
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print('Accuracy:', accuracy)
```

```
↳ Accuracy: 0.7940298507462686
```

LogisticRegression

```
▶ from sklearn.linear_model import LogisticRegression
  # Create an instance of LogisticRegression
  model = LogisticRegression()

  # Train the logistic regression model
  model.fit(X_train, y_train)

  # Make predictions on the test set using logistic regression
  y_pred = model.predict(X_test)

  # Calculate accuracy for logistic regression
  accuracy = accuracy_score(y_test, y_pred)
  print('Accuracy (Logistic Regression):', accuracy)
```

```
↳ Accuracy (Logistic Regression): 0.6865671641791045
```


XGB Classifier



```
# Create an instance of XGBClassifier
model5 = xgb.XGBClassifier()

# Train the model
model5.fit(X_train, y_train)

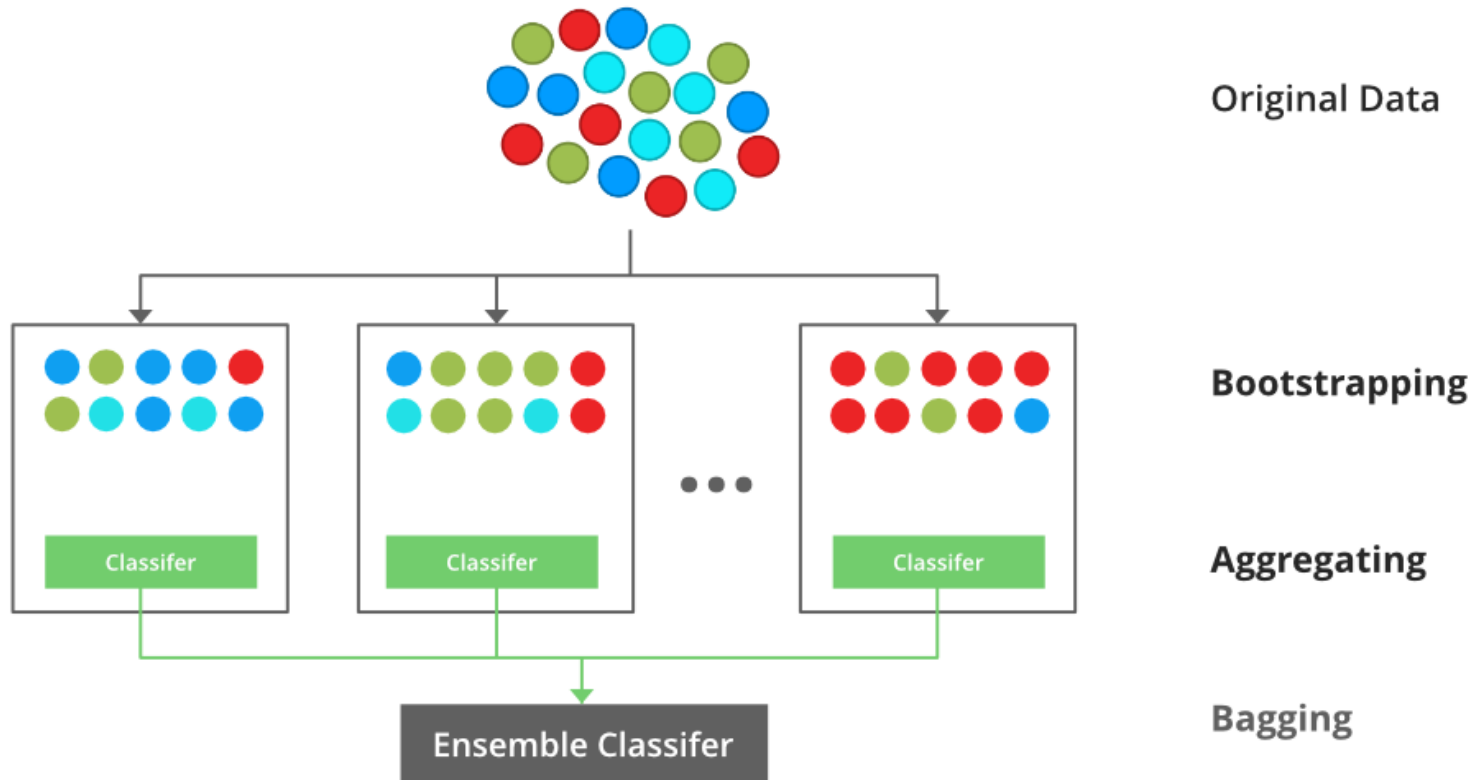
# Make predictions on the test set
y_pred = model5.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print('Accuracy:', accuracy)
```



Accuracy: 0.8

About the M.L Model(XGB Classifier)



- XGBClassifier uses an ensemble of weak prediction models called decision trees. It builds trees sequentially, where each subsequent tree tries to correct the mistakes made by the previous trees. This process is known as gradient boosting.
- XGBClassifier optimizes an objective function that quantifies the model's performance. The objective function measures the difference between the predicted and actual values and guides the algorithm to minimize this difference.

Deploy the Model

To deploy the model first we need to load the model as a file(.jkl) for which we are using the joblib library function.

Save the model for further use

```
import joblib

# Save the trained model to a file
joblib.dump(model5, 'xgb_model.pkl')


['xgb_model.pkl']

[21] # Load the saved Random Forest Classifier
      model5 = joblib.load('xgb_model.pkl')
```

- After saving the model we need to get the user input.
- Then deploy the model with the user input data.

Deploying the Machine Learning Model

Where the user input is collected in the `getvalue_`.



```
# Create a DataFrame from user input
#input_data = pd.DataFrame({'Age': [31], 'Gender': ['Female'], 'Department': ['Engineering'],
#'Designation': ['Senior Engineer'], 'Workperiod in Months': [36]})

# Using the user input
input_data = pd.DataFrame({'Age': [getvalue3],
                           'Gender': [getvalue2],
                           'Department': [getvalue1],
                           'Designation': [getvalue],
                           'Workperiod in Months': [getvalue4]})

# Encoding categorical features using one-hot encoding
input_data_encoded = pd.get_dummies(input_data)
input_data_encoded = input_data_encoded.reindex(columns=X_train.columns, fill_value=0)

# Make the prediction using the trained model
retention_prediction = model5.predict(input_data_encoded)
print(retention_prediction)

# Display the predicted retention status
if retention_prediction == [1]:
    print("The employee is predicted to stay.")
else:
    print("The employee is predicted to leave.")
```

[0]

The employee is predicted to leave.

```
# Create a DataFrame from user input

#input_data = pd.DataFrame({'Age': [34], 'Gender': ['Male'], 'Department': ['Engineering'],
# 'Designation': ['Associate Architect'], 'Workperiod in Months': [114]})

# Using the user input
input_data = pd.DataFrame({'Age': [getvalue3],
                           'Gender': [getvalue2],
                           'Department': [getvalue1],
                           'Designation': [getvalue],
                           'Workperiod in Months': [getvalue4]})

# Encoding categorical features using one-hot encoding
input_data_encoded = pd.get_dummies(input_data)
input_data_encoded = input_data_encoded.reindex(columns=X_train.columns, fill_value=0)

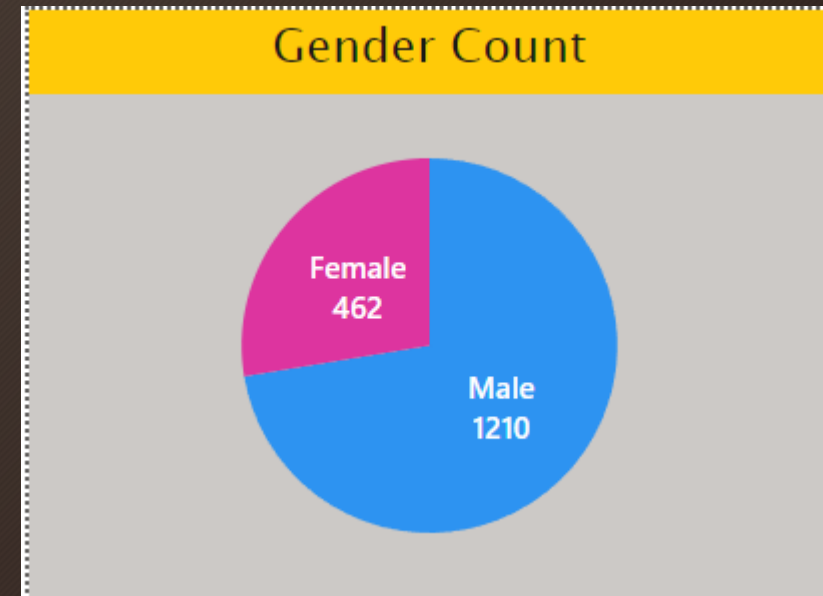
# Make the prediction using the trained model
retention_prediction = model5.predict(input_data_encoded)
print(retention_prediction)

# Display the predicted retention status
if retention_prediction == [1]:
    print("The employee is predicted to stay.")
else:
    print("The employee is predicted to leave.")
```

```
☞ [1]
The employee is predicted to stay.
```


Visualization

- Data visualization is the most important step in the data analysis and prediction process, because the visuals would be easily captured by our brain then the text.
- We have used the Power BI software to visual the dataset.
- And this report is linked with the Power BI visual so we could directly have the interaction.



Employee Report

Total Employee

1672

Total Role

22

Avg Age

32

Department

Engineering

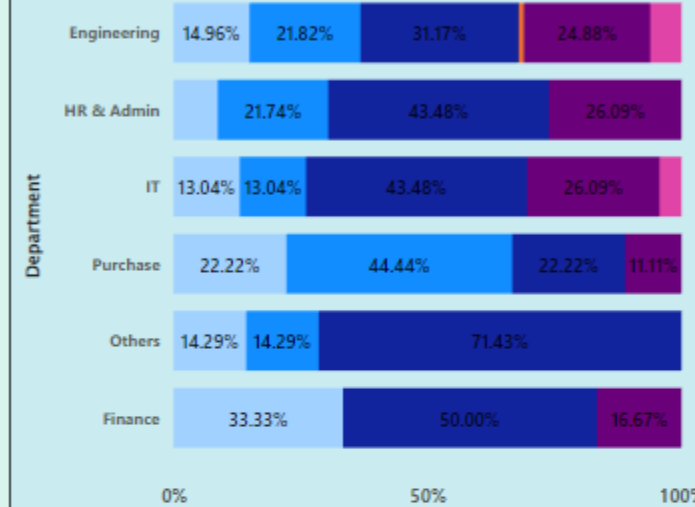
Finance

HR & Admin

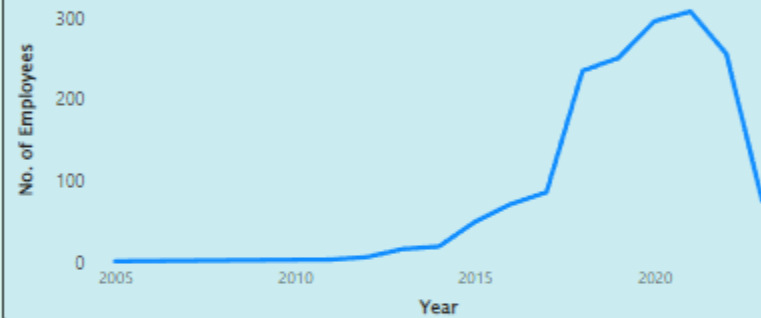


Employees leave based on their Reason

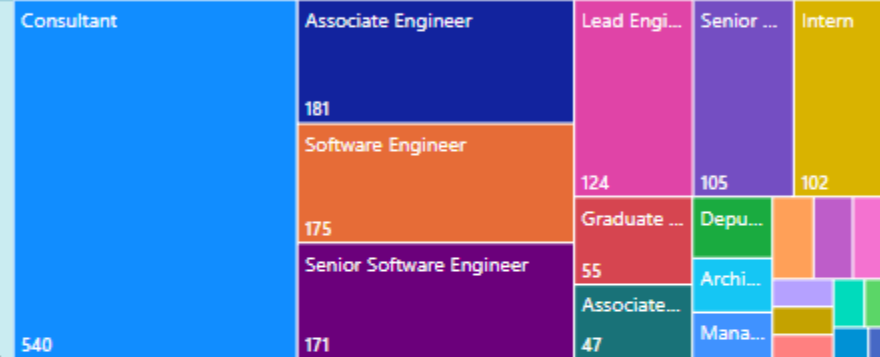
Reason ● Better Opp... ● Career Gr... ● Currently ... ● Higher Ed...



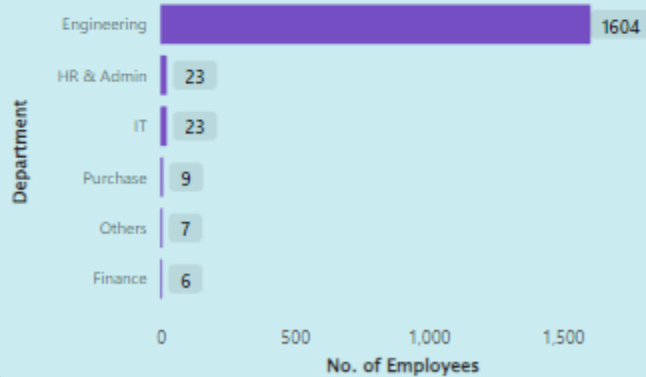
Employees Count Based on their Joining Date



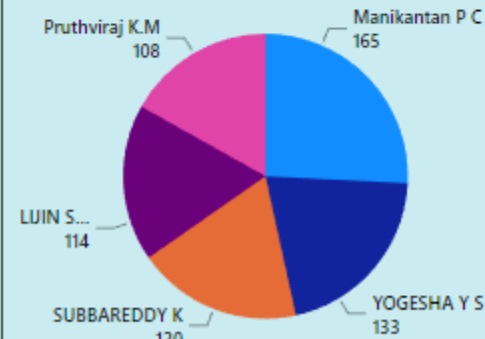
Dept Count by Designation



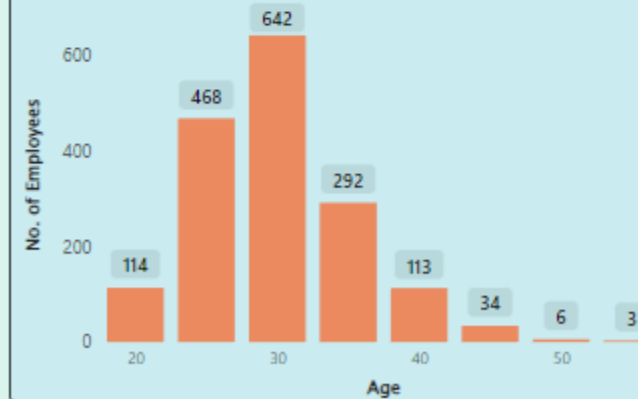
Employees Count Based on their Dept



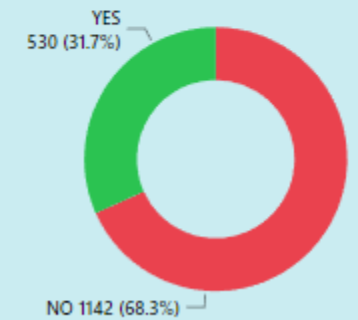
Top 5 Employee by month working



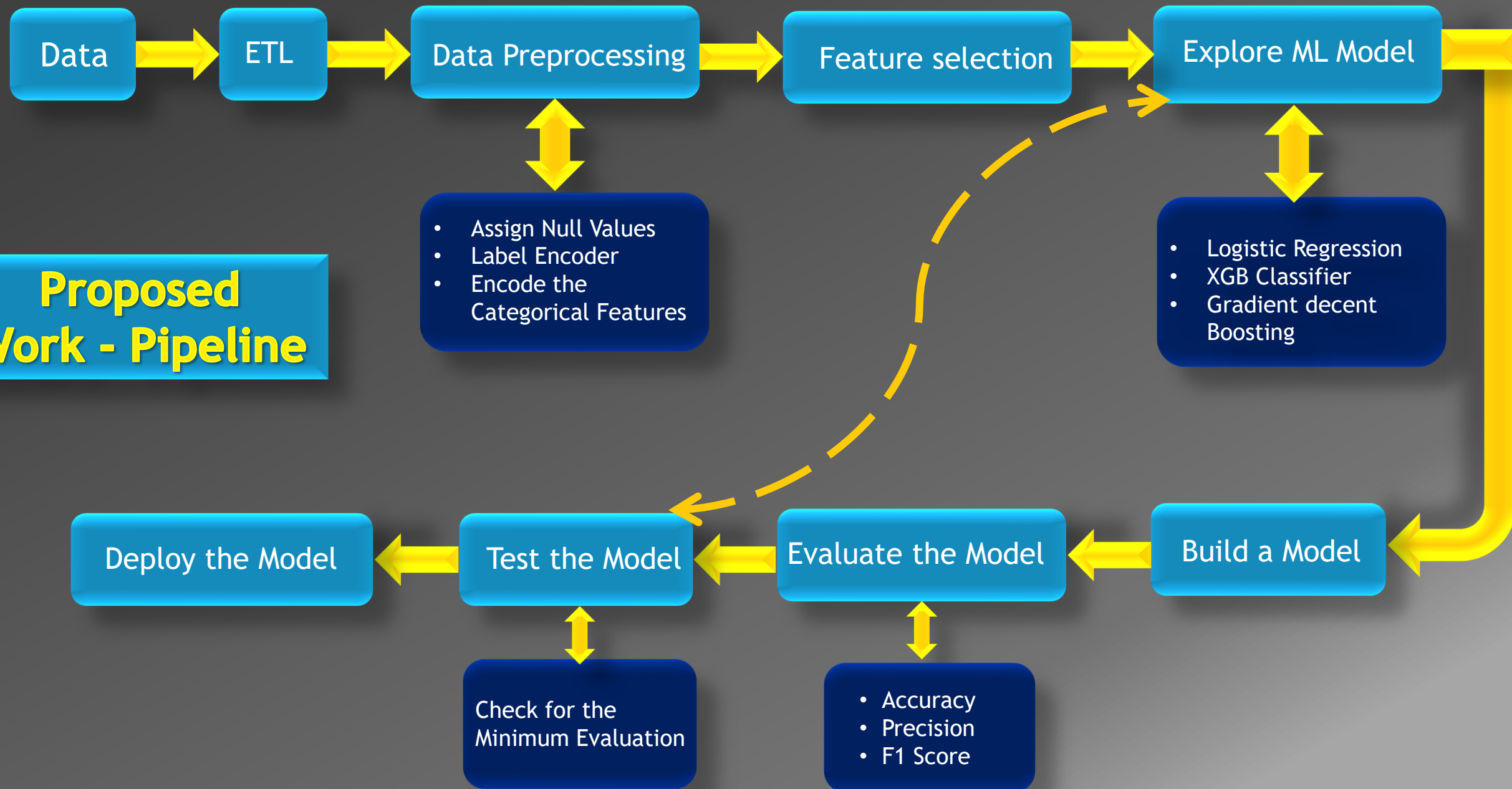
Employees Count Based on their Age



Employees Count Based on their Rention



Proposed Work - Pipeline



Conclusion

- Thus a complete end to end ML pipeline was explored for predicting the employee retention rate.
- The dataset is a good representative of the general workforce in today's organizations. The good
- results from multiple classifiers justify that the features chosen are causes that contribute to voluntary attrition.
- The XGBoost classifier performed well than other ML algorithms with a validation accuracy of 80%
- The reason for attrition of employees can't be exactly predicted, because each person would have different ideas for their future goals.
- Future work might include more number of attributes pertaining to the employee and a Sentiment Analysis can be made by collecting data from employees.

Thank You

- S.Jeevith Kumar