HR ANALYTICS

EMPLOYEE ATTRITION

Team 13

- KEERTHI ANAND
- SUBBAIAH VEERAMANI

BUSINESS PROBLEM

Employees are the backbone of the organization. Organization's performance is heavily based on the quality of the employees.

Challenges that an organization has to face due employee attrition are:

01 **Expensive in terms of both money and time to train new employees.**

02 **Loss of experienced employees**

03 **Impact in productivity**

04 **Impact profit**

BUSINESS QUESTIONS

Having clarity on questions is very crucial because the solution that is being developed will make sense only if we have well stated problem.

01 WHAT FACTORS ARE CONTRIBUTING MORE TO EMPLOYEE ATTRITION?

02 WHAT TYPE OF MEASURES SHOULD THE COMPANY TAKE IN ORDER TO RETAIN THEIR EMPLOYEES?

03 WHAT BUSINESS VALUE DOES THE MODEL BRING?

04 WHICH BUSINESS UNIT FACES THE ATTRITION PROBLEM?

DATASET DESCRIPTION

The following is the IBM Employee churn dataset. It has 35 features with the following datatypes:

Age	int64
Attrition	enum
BusinessTravel	enum
DailyRate	int64
Department	enum
DistanceFromHome	int64
Education	int64
EducationField	enum
EmployeeCount	int64
EmployeeNumber	int64
EnvironmentSatisfaction	int64
Gender	enum
HourlyRate	int64
JobInvolvement	int64
JobLevel	int64
JobRole	enum
JobSatisfaction	int64
MaritalStatus	enum

MonthlyIncome	int64
MonthlyRate	int64
NumCompaniesWorked	int64
Over18	enum
OverTime	enum
PercentSalaryHike	int64
PerformanceRating	int64
RelationshipSatisfaction	int64
StandardHours	int64
StockOptionLevel	int64
TotalWorkingYears	int64
TrainingTimesLastYear	int64
WorkLifeBalance	int64
YearsAtCompany	int64
YearsInCurrentRole	int64
YearsSinceLastPromotion	int64
YearsWithCurrManager	int64

EXPLORATORY DATA ANALYSIS

Supervised learning algorithms are trained on labeled datasets, learning to map inputs to outputs. They are widely used in classification and regression tasks, enabling predictions and decisions based on previous experiences.

01 UNIQUE COUNTS

We notice that 'EmployeeCount', 'Over18', 'StandardHours' have only one unique values and 'EmployeeNumber' has 1470 unique values. This features aren't useful for us, So we are going to drop those columns.

```
Age: Number of unique values 43
==========================================================
Attrition: Number of unique values 2
==========================================================
BusinessTravel: Number of unique values 3
==========================================================
DailyRate: Number of unique values 886
==========================================================
Department: Number of unique values 3
==========================================================
DistanceFromHome: Number of unique values 29
==========================================================
Education: Number of unique values 5
==========================================================
EducationField: Number of unique values 6
==========================================================
EmployeeCount: Number of unique values 1
==========================================================
EmployeeNumber: Number of unique values 1470
==========================================================
```

02 CATEGORICAL COLUMNS

```
====================================
MaritalStatus : ['Married', 'Divorced', 'Single']
+-------------+-----+
|MaritalStatus|count|
+-------------+-----+
|      Married|  673|
|       Single|  470|
|     Divorced|  327|
+-------------+-----+


====================================
OverTime : ['No', 'Yes']
+--------+-----+
|OverTime|count|
+--------+-----+
|      No| 1054|
|     Yes|  416|
+--------+-----+


====================================
```

EXPLORATORY DATA ANALYSIS

03 NUMERICAL FEATURES

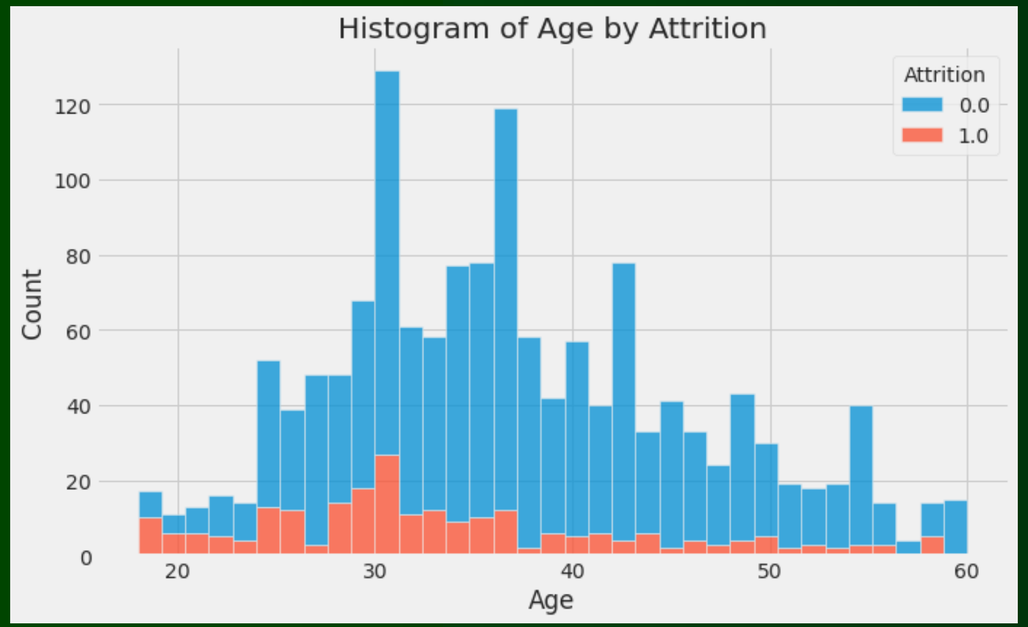
```
Age : Minimum: 18, Maximum: 60
===================================
DailyRate : Minimum: 102, Maximum: 1499
===================================
HourlyRate : Minimum: 30, Maximum: 100
===================================
MonthlyIncome : Minimum: 1009, Maximum: 19999
===================================
MonthlyRate : Minimum: 2094, Maximum: 26999
===================================
TotalWorkingYears : Minimum: 0, Maximum: 40
===================================
YearsAtCompany : Minimum: 0, Maximum: 40
===================================
```

04 MISSING VALUES

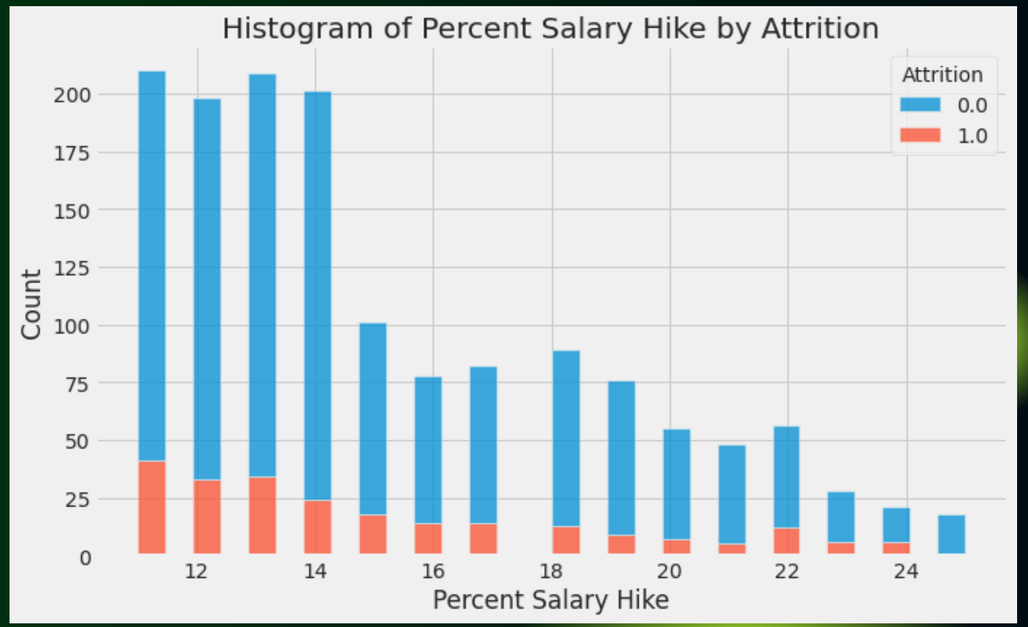
Label	Type	Missing	Zeros	PosInf	NegInf	Min	Max
<fct>	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Age	int	0	0	0	0	18	60
Attrition	enum	0	874	0	0	0	1
BusinessTravel	enum	0	105	0	0	0	2
DailyRate	int	0	0	0	0	102	1499
Department	enum	0	49	0	0	0	2
DistanceFromHome	int	0	0	0	0	1	29
Education	int	0	0	0	0	1	5
EducationField	enum	0	21	0	0	0	5
EmployeeCount	int	0	0	0	0	1	1
EmployeeNumber	int	0	0	0	0	1	2065
EnvironmentSatisfaction	int	0	0	0	0	1	4
Gender	enum	0	427	0	0	0	1
HourlyRate	int	0	0	0	0	30	100
JobInvolvement	int	0	0	0	0	1	4

EXPLORATORY DATA ANALYSIS

05 AGE

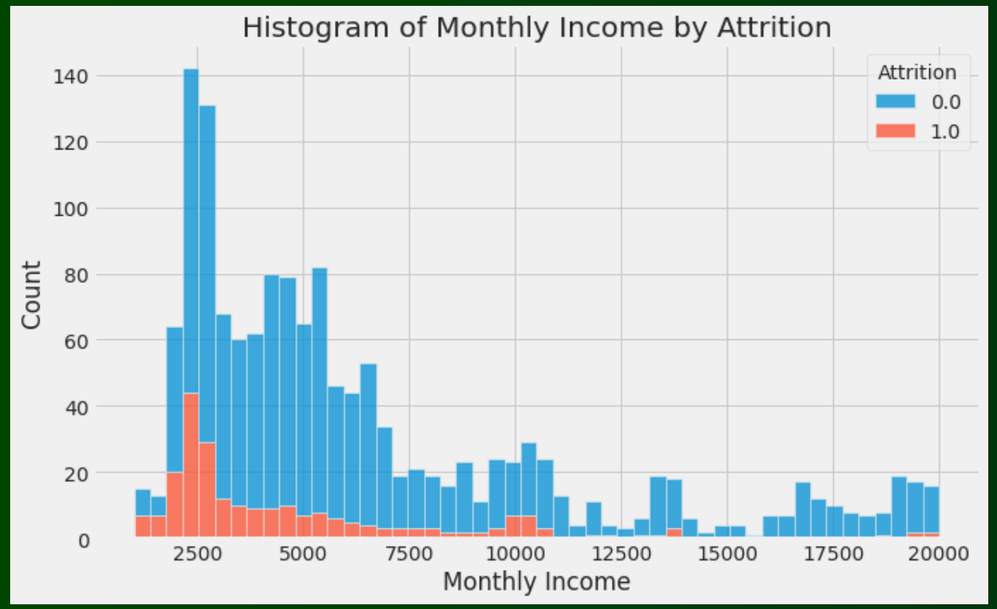


06 SALARY HIKE

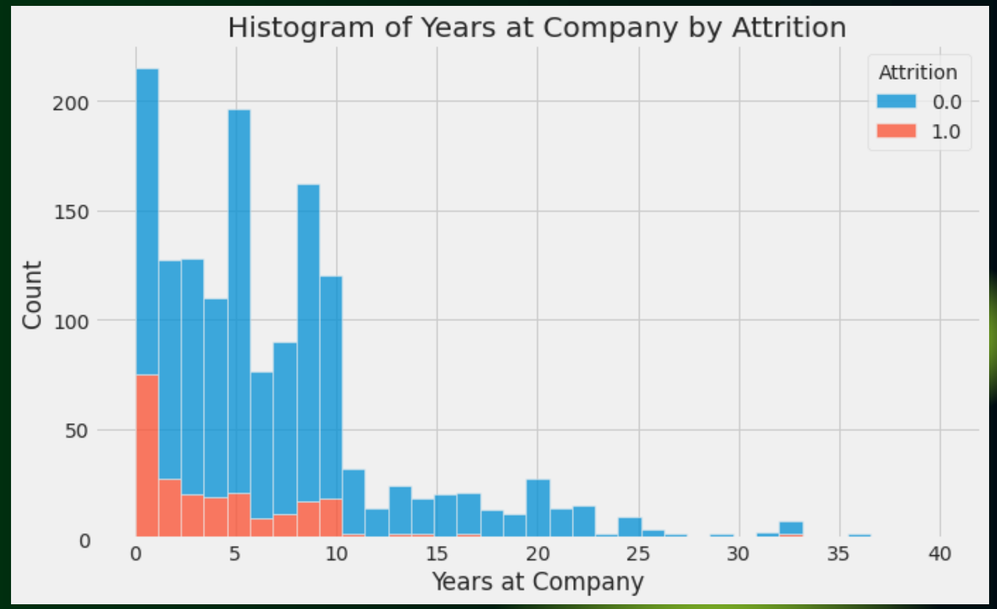


EXPLORATORY DATA ANALYSIS

07 MONTHLY INCOME



08 YEARS AT COMPANY



EDA CONCLUSIONS

- The workers with low **JobLevel, MonthlyIncome, YearAtCompany, and TotalWorkingYears** are more likely to quit there jobs.
- **BusinessTravel :** The workers who travel alot are more likely to quit then other employees.
- **Department :** The worker in Research & Development are more likely to stay then the workers on other departement.
- **EducationField :** The workers with Human Resources and Technical Degree are more likely to quit then employees from other fields of educations.
- **Gender :** The Male are more likely to quit.
- **JobRole :** The workers in Laboratory Technician, Sales Representative, and Human Resources are more likely to quit the workers in other positions.
- **MaritalStatus :** The workers who have Single marital status are more likely to quit the Married, and Divorced.
- **OverTime :** The workers who work more hours are likely to quit then others.

SPARK SQL QUERIES:

We have found some insights by the following Spark SQL queries

01 What is the attrition rate in the organization?

attrition_rate
16.12244898

02 WHICH DEPARTMENT HAS THE HIGHEST ATTRITION RATE?

Department	attrition_rate
Sales	20.62780269
Human Resources	19.04761905
Research & Development	13.83975026

03 WHICH BUSINESS UNIT IS LOSING ITS HIGH PERFORMERS?

Department	PerformanceRating	high_performer_attrition_count	total_high_performers	high_performer_attrition_rate
Research & Development	4	26	156	16.66666667
Sales	4	10	61	16.39344262
Human Resources	4	1	9	11.11111111

SPARK SQL QUERIES:

04 What is the impact of business travel on attrition?

BusinessTravel	attrition_rate
Travel_Frequently	24.90974729
Travel_Rarely	14.95685523
Non-Travel	8

05 WHAT JOB ROLES HAVE THE HIGHEST ATTRITION?

JobRole	attrition_rate
Sales Representative	39.75903614
Laboratory Technician	23.93822394
Human Resources	23.07692308
Sales Executive	17.48466258
Research Scientist	16.09589041

06 HOW DOES A COMBINATION OF JOB ROLE, OVERTIME, AND JOB SATISFACTION IMPACT ATTRITION?

JobRole	OverTime	JobSatisfaction	attrition_count	total_employees	attrition_rate
Sales Representative	Yes	2	7	7	100
Human Resources	Yes	1	3	3	100
Laboratory Technician	Yes	2	3	4	75
Sales Representative	No	1	7	11	63.63636364
Sales Representative	Yes	4	4	7	57.14285714
Laboratory Technician	Yes	3	10	18	55.55555556
Sales Representative	Yes	3	5	9	55.55555556

PREDICTIVE MODELING USING SPARK MLLIB

01 OBJECTIVE

- Develop machine learning models to predict employee attrition.
- Compare performance of different classification models using Spark MLlib.

02 DATA PREPROCESSING FOR ML

- **Categorical Encoding:** Used **StringIndexer** & **OneHotEncoder** to transform categorical features.
- **Feature Engineering:** Used **VectorAssembler** to create a feature vector for ML models.
- **Train-Test Split:** Split data into 80% training & 20% test set.

PREDICTIVE MODELING USING SPARK MLLIB

03 MODELS TRAINED

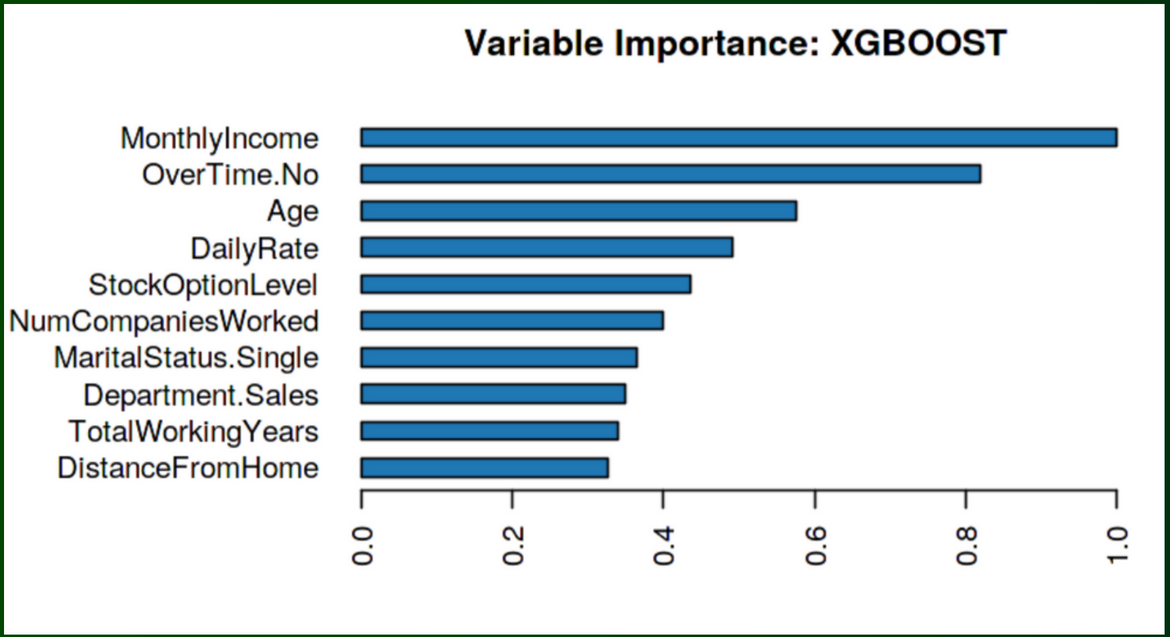
We implemented four classification models to predict attrition:

- **Logistic Regression –** Baseline model for benchmarking.
- **Random Forest –** Strong tree-based model for structured data.
- **XGBoost –** Advanced boosting algorithm, highly effective for tabular data.
- **CatBoost –** Best suited for datasets with categorical-heavy features.

Pipeline Approach: Automated feature transformations & model training.

PREDICTIVE MODELING USING SPARK MLLIB

04 FEATURE IMPORTANCE



Top Reasons why Employees leave the Organization:

- No Overtime
- Monthly Income
- Age

PREDICTIVE MODELING USING SPARK MLLIB

05 MODEL PERFORMANCE COMPARISON

Metrics Evaluated:

- **Accuracy** → Measures overall correctness.
- **F1-Score** → Harmonic mean of precision and recall, useful for imbalanced datasets.
- **AUC (Area Under ROC Curve)** → Assesses model discrimination power.
- **Precision-Recall Score** → Evaluates model's effectiveness for imbalanced data.

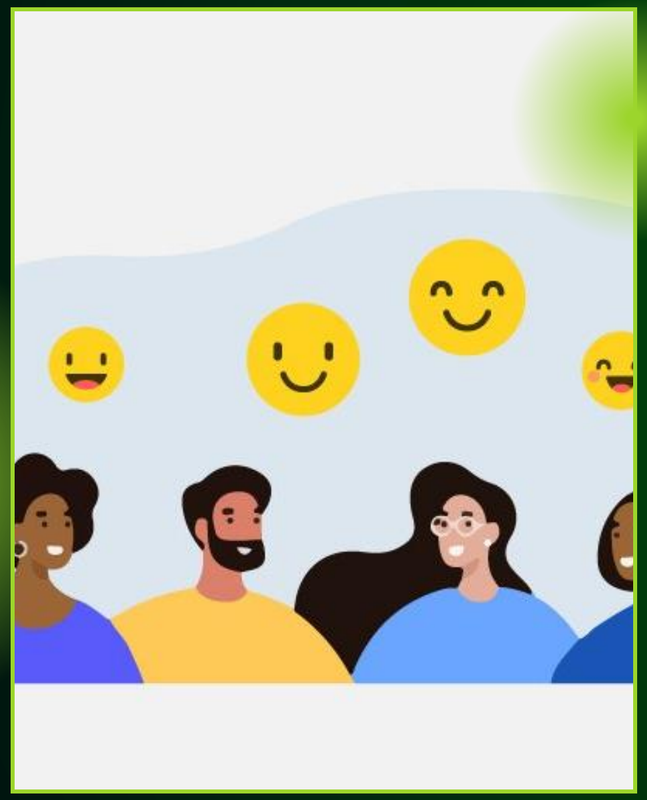
MODEL	ACCURACY	F1- SCORE	AUC (ROC)	Precision-Recall
Logistic Regression	86.22%	0.8581	0.8114	0.5492
Random Forest	85.43%	0.8036	0.7929	0.5081
XGBoost	87.01%	0.3187	0.8003	0.5155
CatBoost	88.98%	0.3764	0.8227	0.6330

PREDICTIVE MODELING USING SPARK MLLIB

FINAL TAKEAWAYS

- **Best Accuracy: CatBoost (88.98%)** performed best in overall accuracy, indicating strong predictive power.
- **Highest AUC (ROC): CatBoost (0.8227)** showed the best ability to differentiate between employees staying and leaving.
- **Best Precision-Recall AUC: CatBoost (0.6330)** is the most effective at detecting employees likely to leave, making it the best model for HR to take action on high-risk employees.
- **F1-Score Analysis: Logistic Regression (0.8581)** had the highest F1-score, showing better balance between precision & recall, but lacks interpretability compared to tree-based models.
- **Random Forest & XGBoost:** Decent performance, but not as strong as CatBoost, likely due to categorical-heavy dataset where CatBoost excels.

CONCLUSION AND INFERENCES:



01 AGE

People are tending to switch to a different jobs at the start of their careers, or at the earlier parts of it. Once they have settled with a family or have found stability in their jobs, they tend to stay long in the same organization- only going for vertical movements in the same organization.

02 INCENTIVES

Salary and stock options have a great motivation on the employees and people tend to leave the organization much lesser. Higher pay and more stock options have seen more employees remain loyal to their company.

03 WORK-LIFE BALANCE

Work life balance is a great motivation factor for the employees. However, people with a good work-life balance, tend to switch in search of better opportunities and a better standard of living.

04 WORK STRESS

Departments where target meeting performance is very much crucial (for e.g. Sales) tend to have a greater chances of leaving the organization as compared to departments with more administration perspective (For e.g. Human Resources)