Employee Churn Prediction

This presentation outlines the process of predicting employee churn using various machine learning models. We will cover data collection, wrangling, analysis, and modeling, concluding with a comparison of model accuracies.

Team Members

Menna-allah Abdel-Maarouf Ahmed

Razan Ihab Abdellatif

Abdelrahman Ahmed Hussein

Mohamed Khaled Mondy

Data Collection & Representation

Data Representation

A quick look at the first few rows of the DataFrame reveals the structure and types of data collected for each employee.

Salary_INR	Satisfaction	Evaluation	number_of_projects	average_montly_hours	time_spent_compa	work_accident	Promotion	Department
11053	3.8	5.3	2	157	1	1	0	sales
54063	8.0	8.6	5	262	2	0	0	sales
61300	1.1	8.8	7	272	2	0	0	hr
20647	7.2	8.7	5	223	3	1	0	sales
17808	3.7	5.2	2	159	4	0	0	sales

Data Collection & Representation

Key Employee Metrics

Satisfaction: Score from 0.9 to 9.2, indicating overall job contentment.

Evaluation: Performance score ranging from 4.5 to 10.0.

Number of Projects: Assigned projects, between 2 and 7.

Work-Life Balance & History

Average Monthly Hours: Worked between 126 and 310 hours.

Time Spent Company: Tenure at the company, from 2 to 6 years.

Work Accident: Binary indicator (0 for no, 1 for yes).

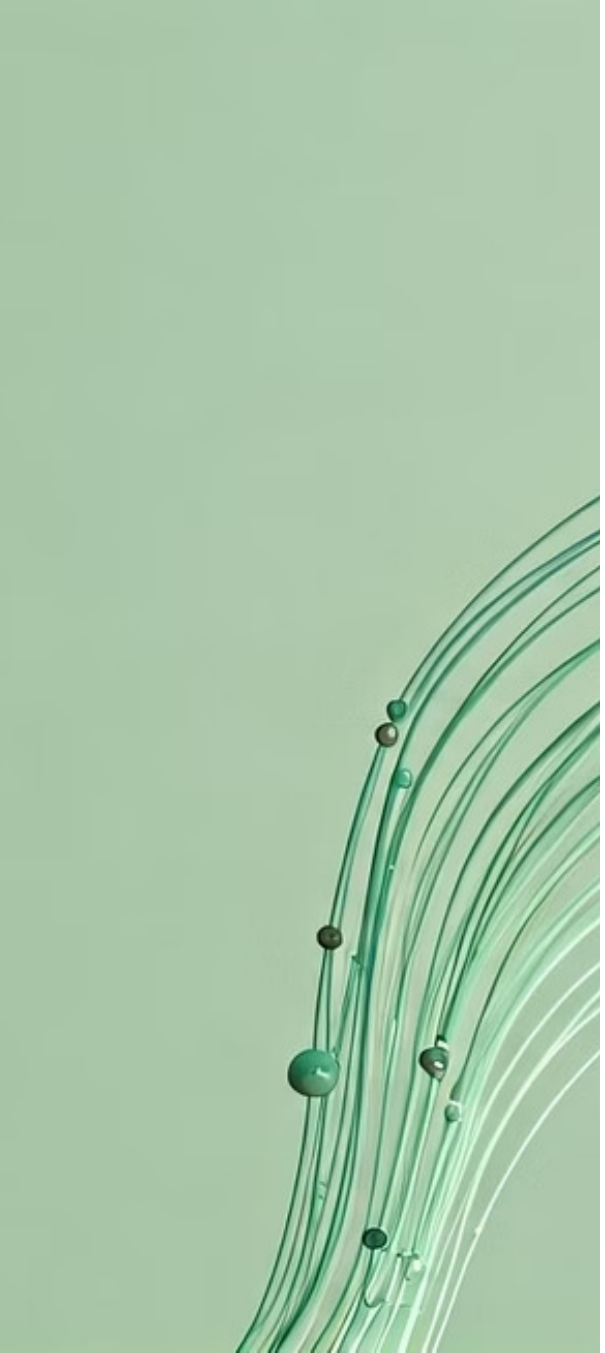
Promotion: Binary indicator (0 for no, 1 for yes).

Departmental & Salary Insights

Department: Categorical data including Sales, Accounting, HR, etc.

Salary (INR): Employee compensation in Indian Rupees.

Churn: The target variable, indicating if an employee left (1) or stayed (0).



Data Wrangling & Analysis Overview

Data Wrangling

Before analysis, data was cleaned by checking for and handling missing values and duplicated entries. No missing values or duplicates were found, ensuring data integrity.

```
df.isna().sum()
```

```
df.duplicated().sum()
```

Data Analysis

An overview of the dataset's information, including data types and non-null counts, confirmed 14,999 entries across 11 columns, with no missing values.

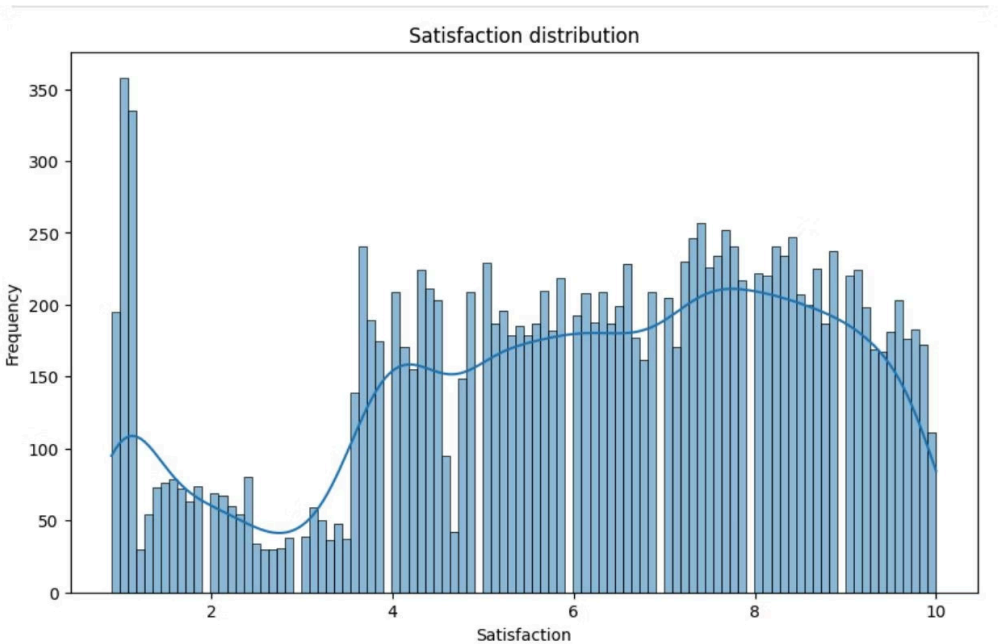
```
df.info()
```

This foundational step prepares the data for deeper statistical examination and visualization.

Key Employee Metrics: Satisfaction & Evaluation

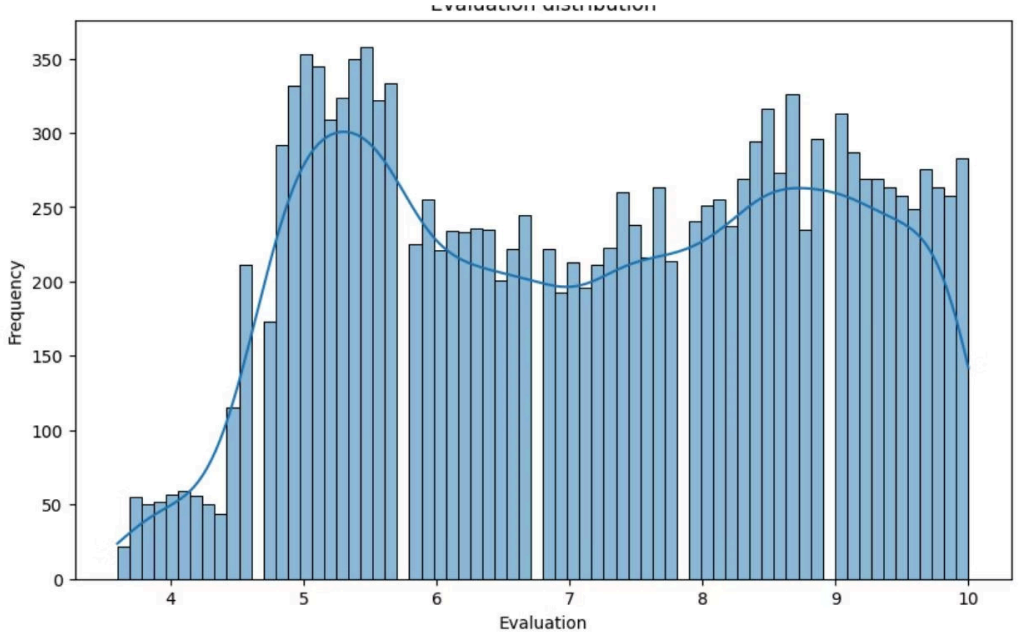
Satisfaction Distribution

Employee satisfaction ranges from 0.9 to 10.0, with a mean of 6.13. The distribution shows a varied sentiment among employees.



Evaluation Distribution

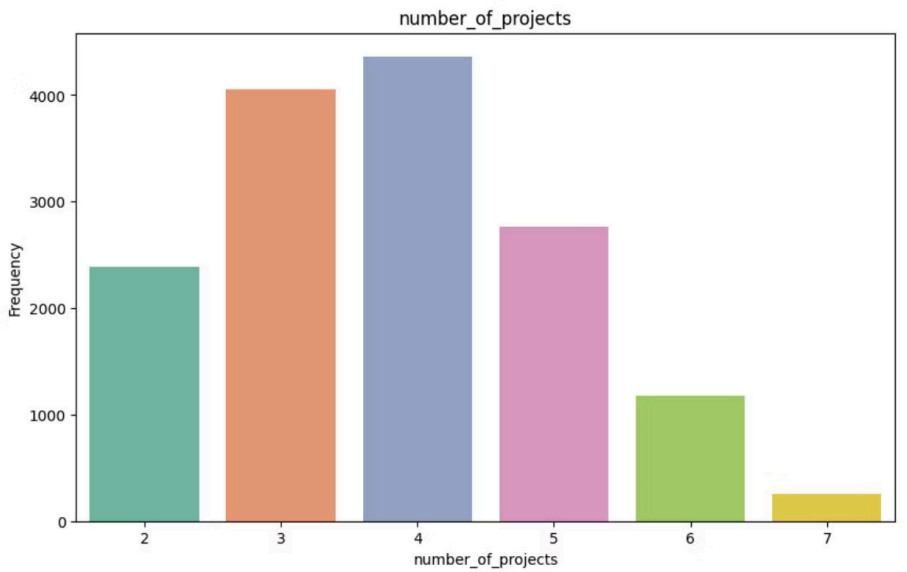
Employee evaluation scores range from 3.6 to 10.0, averaging 7.16. This metric indicates performance levels across the workforce.



Project Load & Time Spent in Company

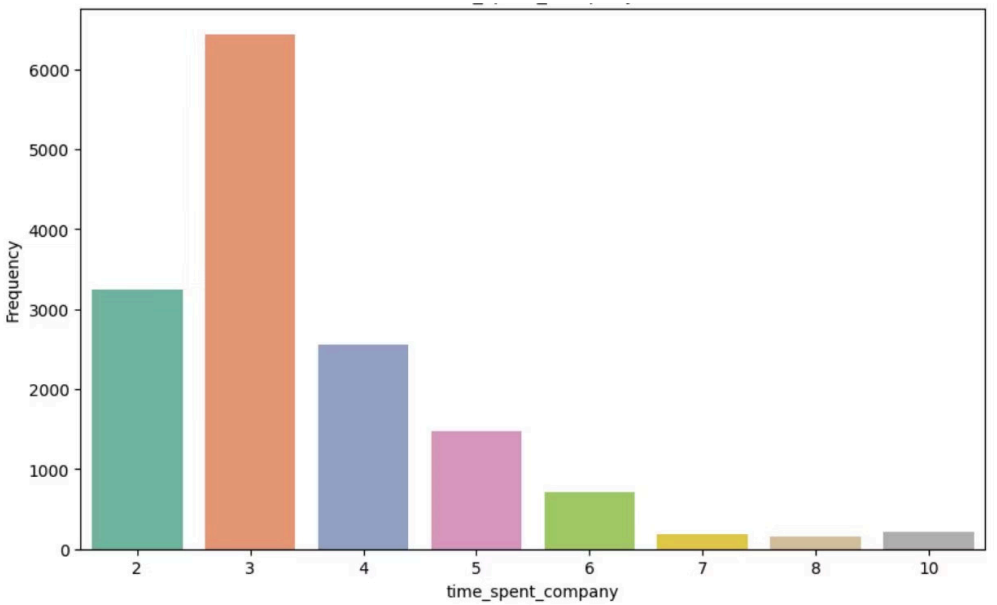
Number of Projects

Most employees are involved in 3 or 4 projects, with fewer working on 6 or 7. This distribution highlights typical workload patterns.



Time Spent in Company

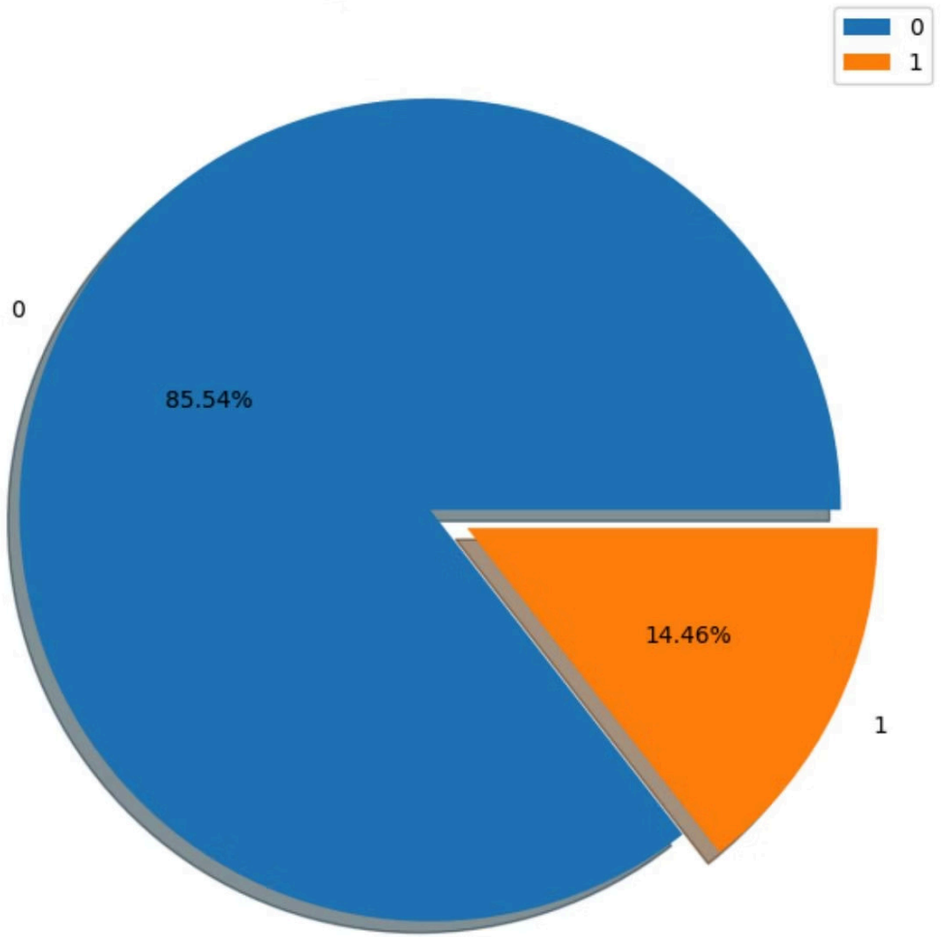
A significant portion of employees have been with the company for 3 years, followed by 2 and 4 years. Fewer stay for 7, 8, or 10 years.



Work Accidents & Promotions

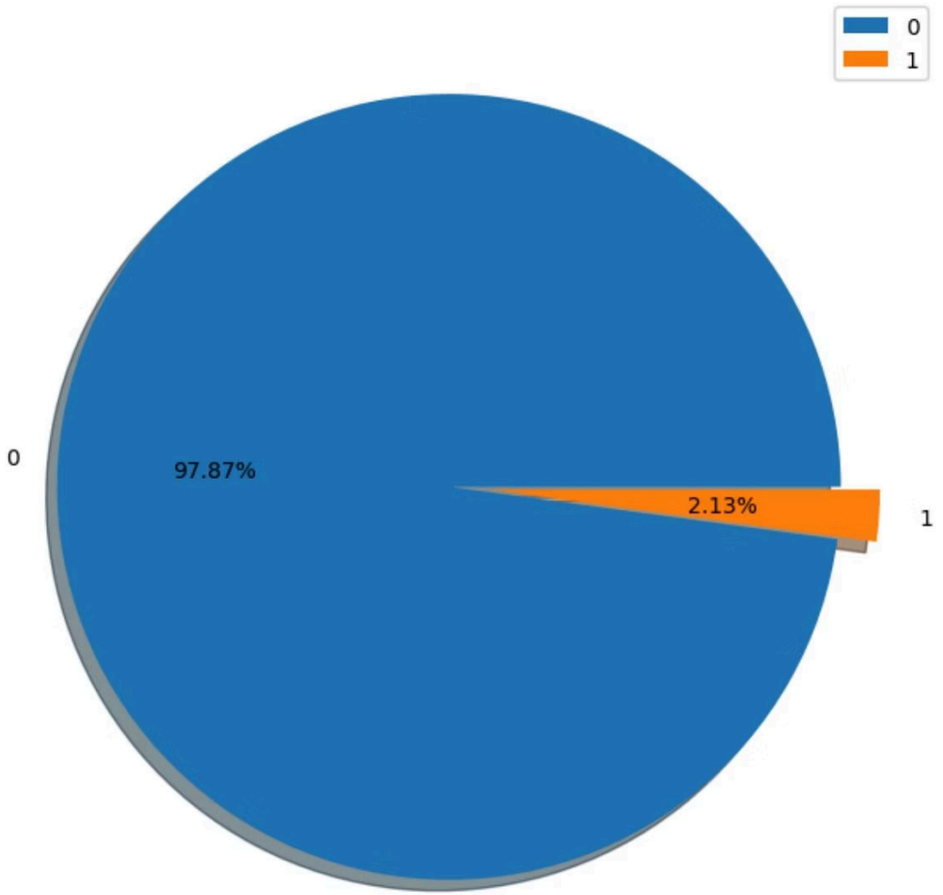
Work Accident Distribution

The data shows that a large majority of employees (85.54%) have not experienced a work accident, while 14.46% have.



Promotion Distribution

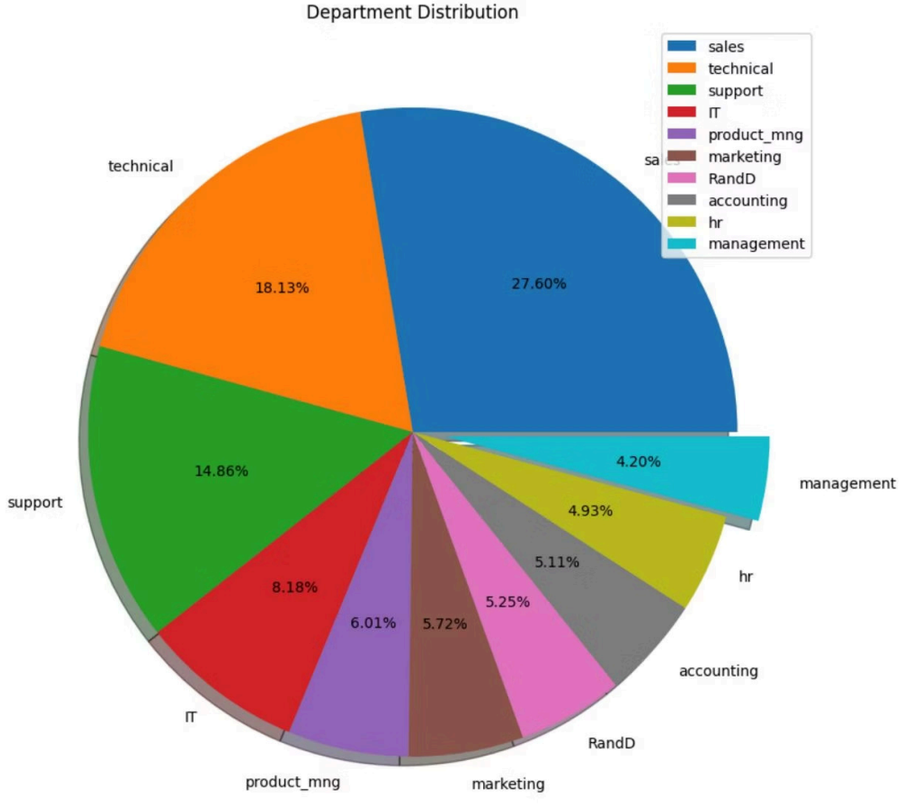
Only a small percentage of employees (2.13%) have received a promotion, indicating a low promotion rate within the company.



Departmental & Salary Insights

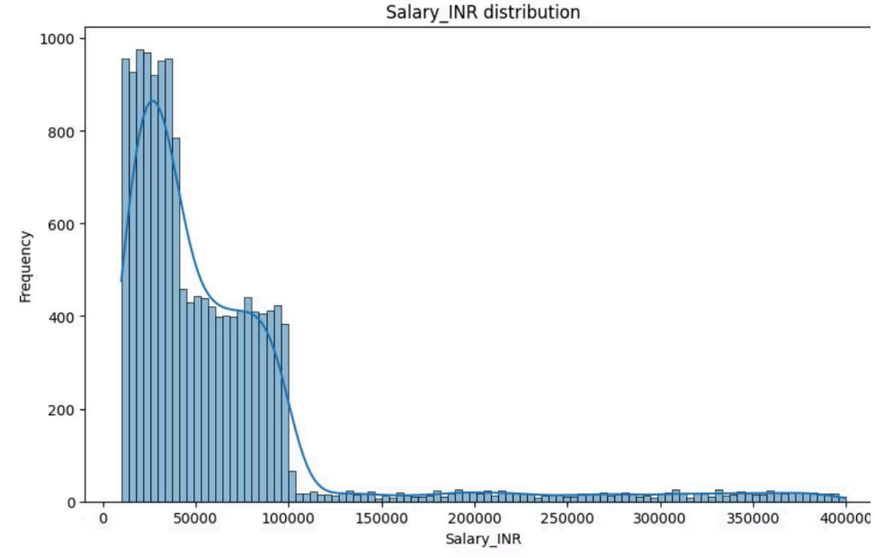
Department Distribution

Sales, technical, and support departments account for the largest portions of the workforce, reflecting the company's operational structure.

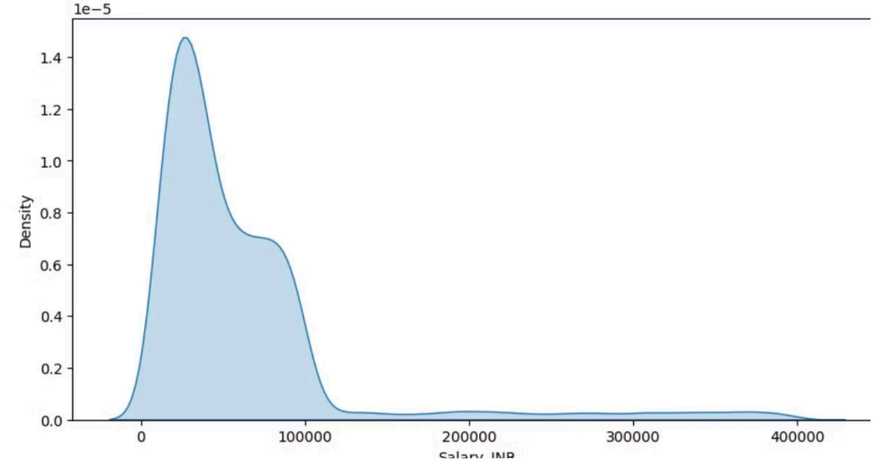


Salary Distribution

Salaries range from INR 10,000 to INR 399,930, with a mean of INR 62,742. This wide range suggests diverse compensation levels.



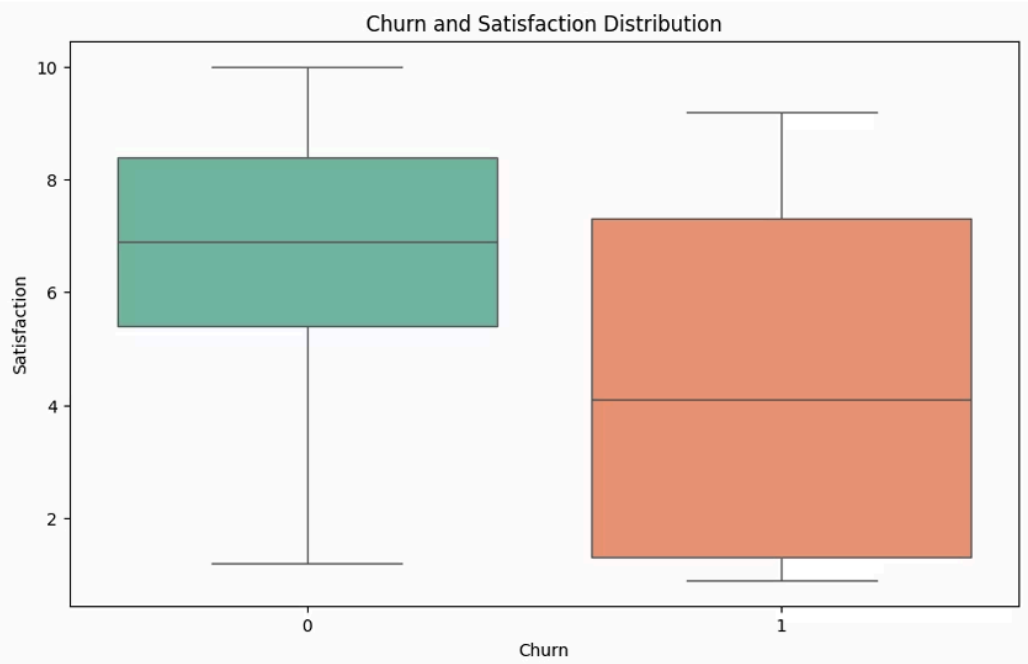

```
In [53]: plt.figure(figsize=(10,5))
         sns.kdeplot(df['Salary_INR'],fill=True)
         plt.show()
```



Exploratory Data Analysis (EDA)

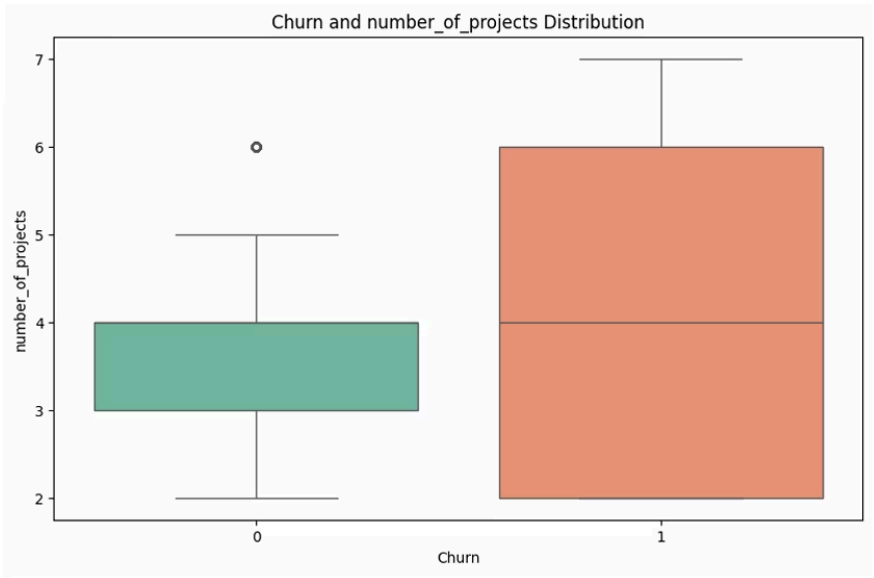
EDA revealed key relationships between employee attributes and churn. Visualizations highlight how satisfaction, evaluation, and other factors correlate with an employee's decision to leave.

Satisfaction vs. Churn



Lower satisfaction levels are strongly associated with higher churn rates.

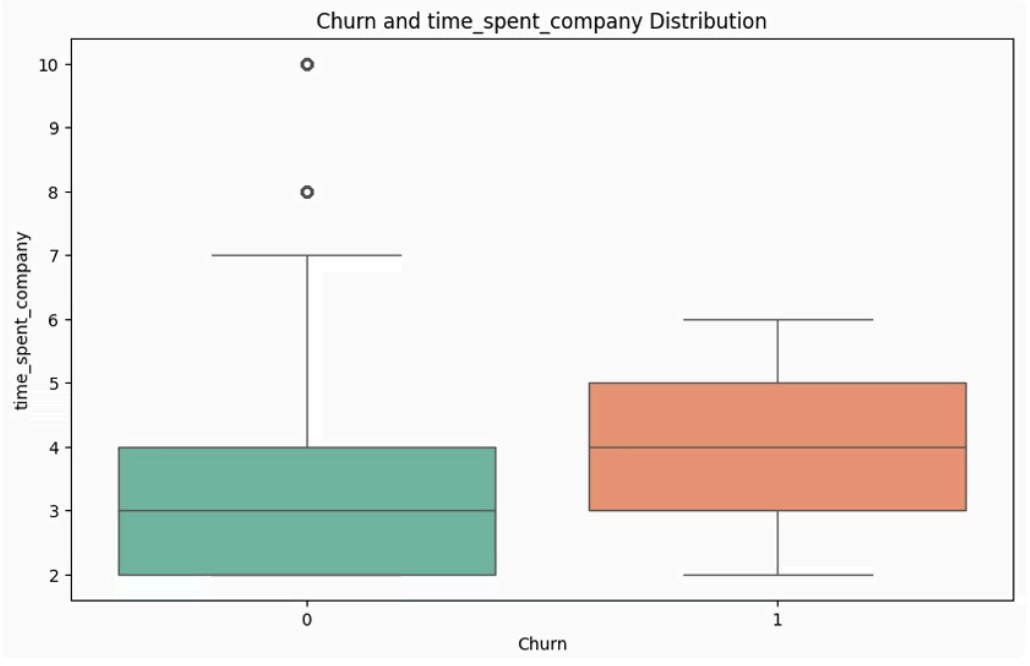
Number Of Projects vs. Churn



Assigning 3 to 5 projects per employee helps minimize churn and maintain productivity.

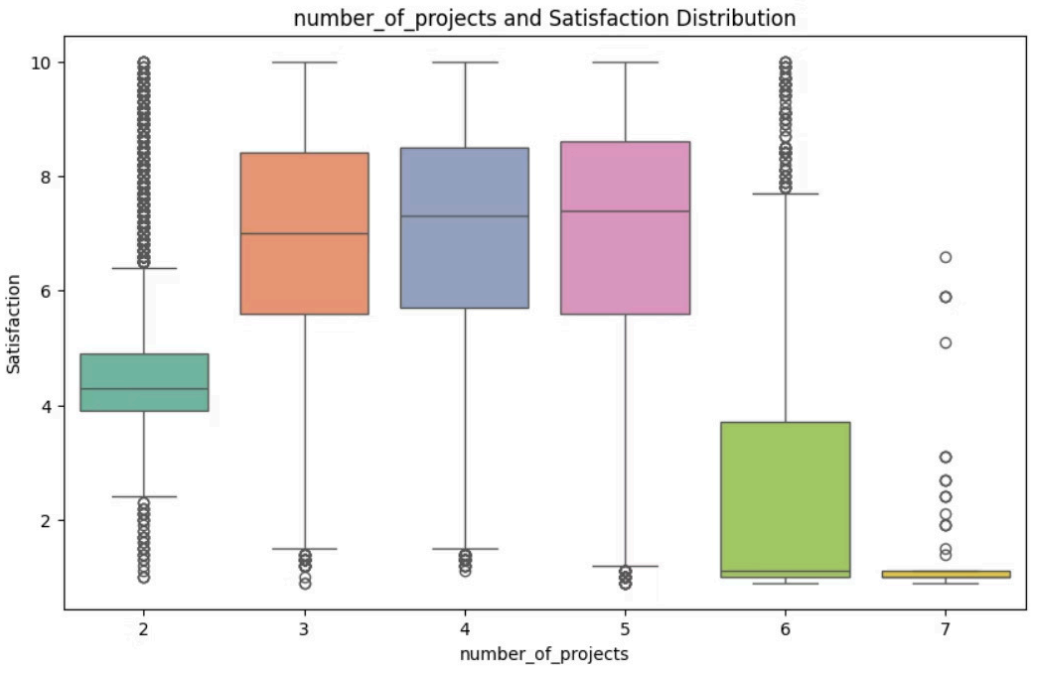
Exploratory Data Analysis (EDA)

Time Spent Company vs. Churn



Analysis reveals that employees who spent **3 to 5 years** at the company show the **highest churn rate**. This suggests that if career growth or promotions are not offered during this period, employees are more likely to leave.

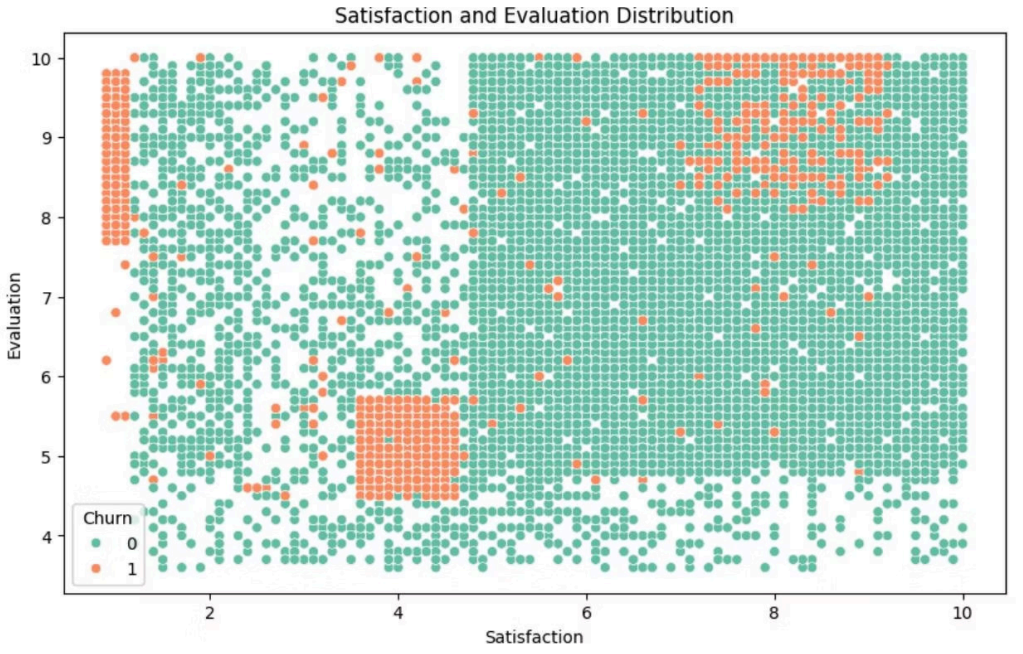
Satisfaction vs. number_of_projects



Employees with **too few** or **too many projects** tend to have **lower satisfaction**. The highest satisfaction levels are seen in those handling a **moderate workload (3–5 projects)**.

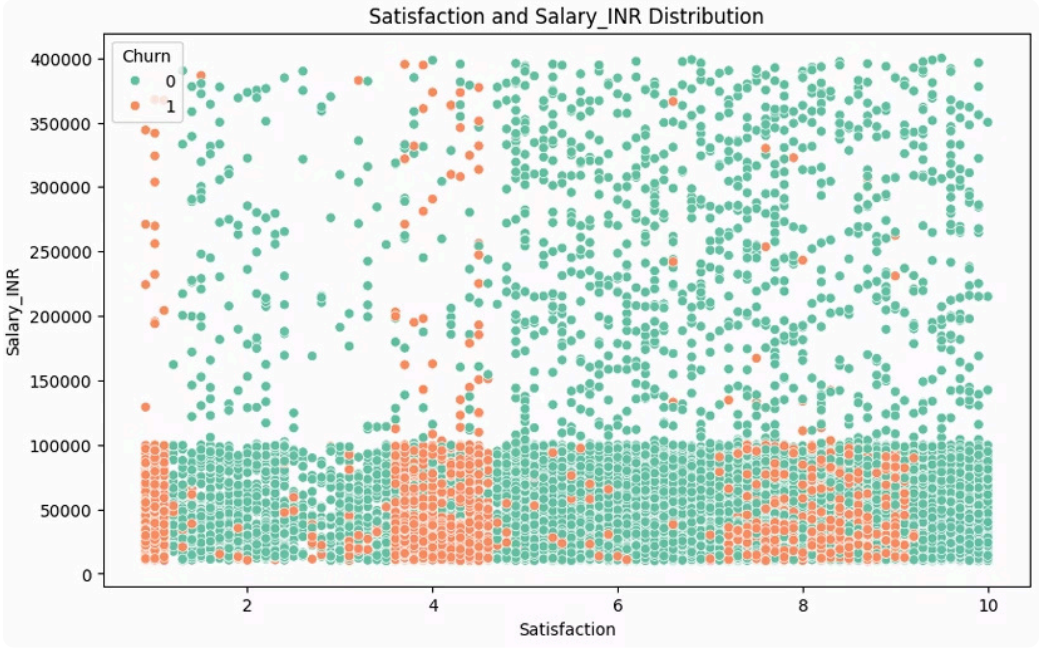
Key Churn Indicators

Satisfaction & Evaluation



Employees with low satisfaction and high evaluation, or low satisfaction and low evaluation, are more likely to churn.

Satisfaction & Salary



There is **no strong correlation** between salary and satisfaction. Some **high-paid employees are still dissatisfied**, while **some low-paid employees are satisfied**.

Key Churn Indicators

Number of Projects & Salary

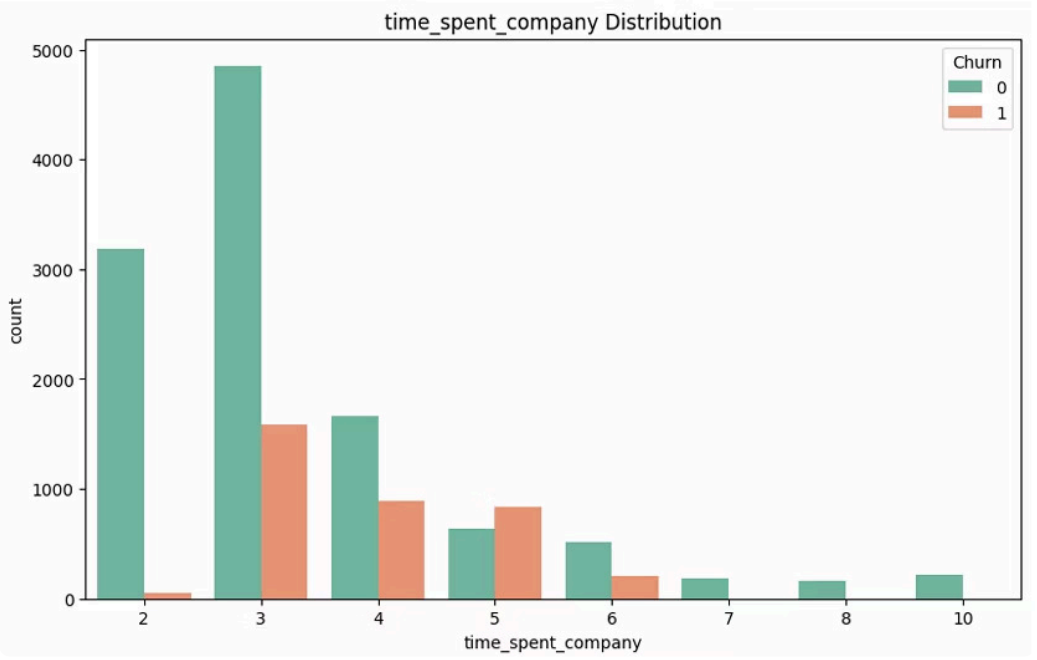
Employees with **lower salaries** are often assigned **more projects**, which may lead to dissatisfaction and higher churn.

Time spent & Salary

Longer time at the company **does not always mean a higher salary**. Some employees with **5+ years** still receive **low pay**, which may lead to dissatisfaction and churn.

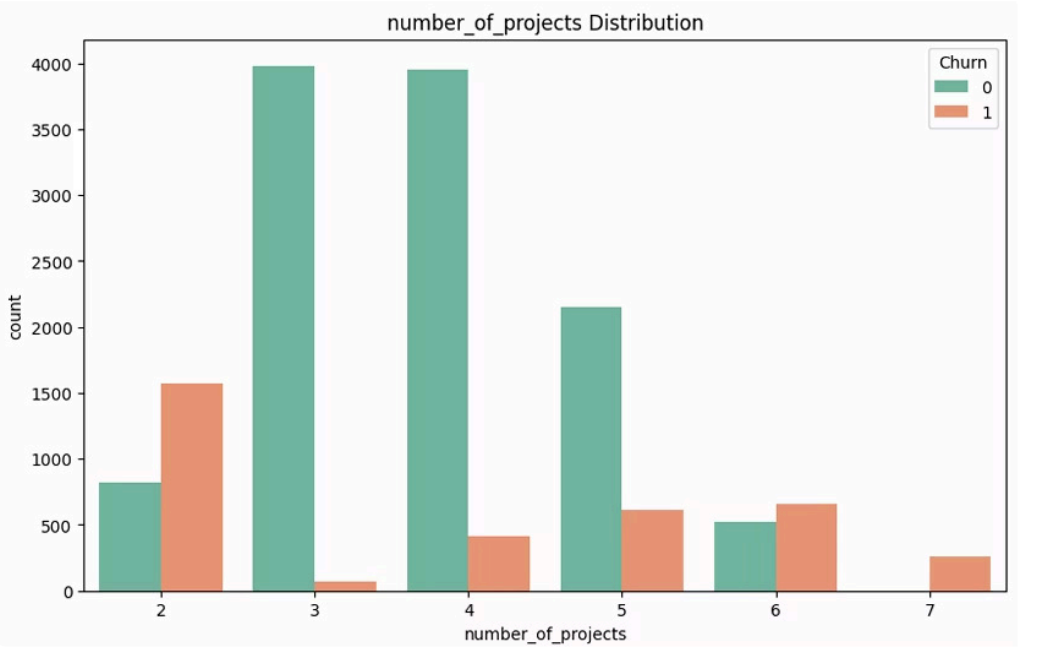
Key Churn Indicators

Time Spent In Company Churn



The majority of employees have spent **3 years** at the company, followed by 2 and 4 years. After **5+ years**, the number of employees drops significantly, but **churn rates increase**.

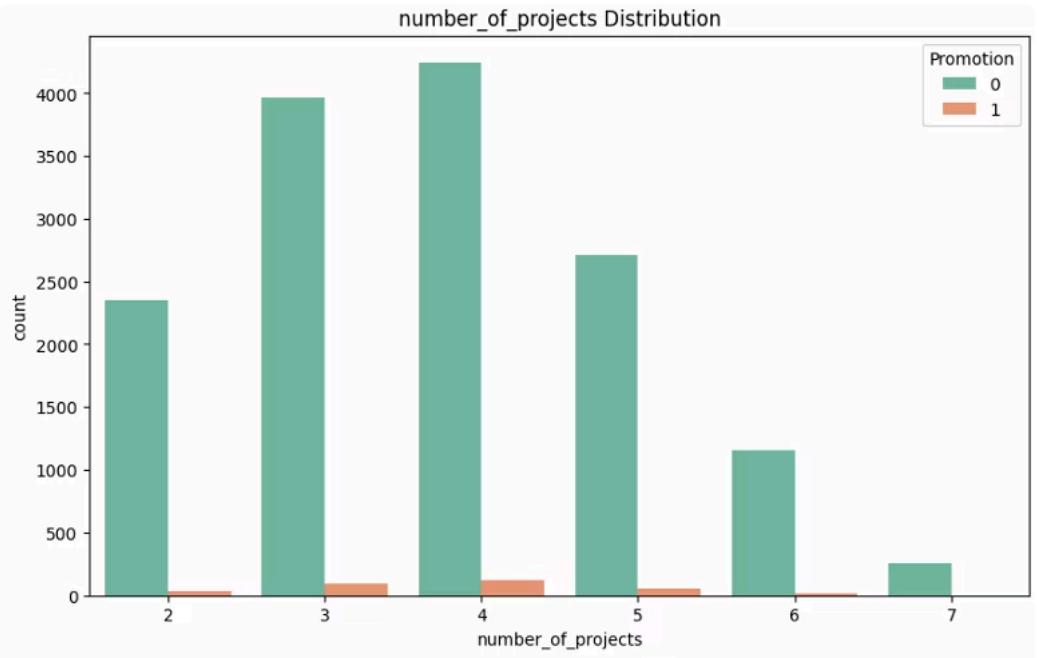
Number of Projects Churn



Employees with **very few (2)** or **too many (6 or 7)** projects have a **higher churn rate**, while those with **3 to 5 projects** show the **lowest churn**.

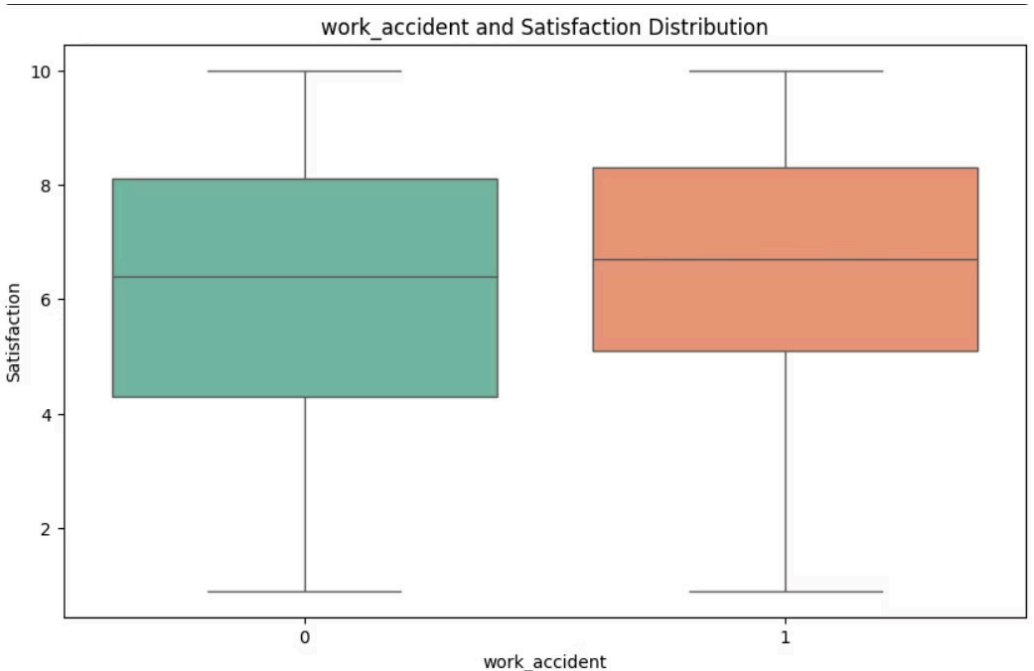
Key Churn Indicators

Time Spent In Company Promotion



The majority of employees have spent **3 years** at the company, followed by 2 and 4 years. After **5+ years**, the number of employees drops significantly, but **churn rates increase**.

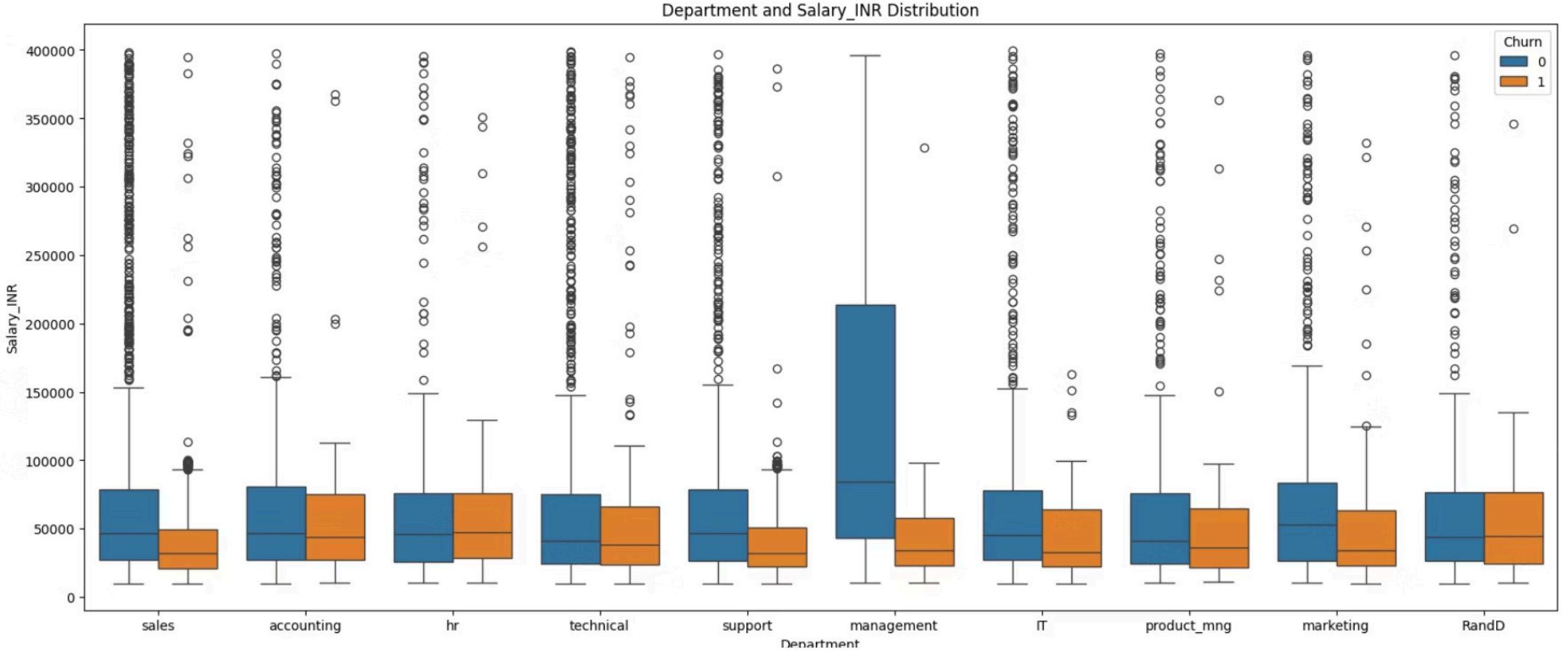
work accident and Satisfaction



Employees who experienced a **work accident** tend to report **slightly lower satisfaction levels** compared to others. Although not the strongest churn driver, it may impact morale and perception of workplace safety.

Key Churn Indicators

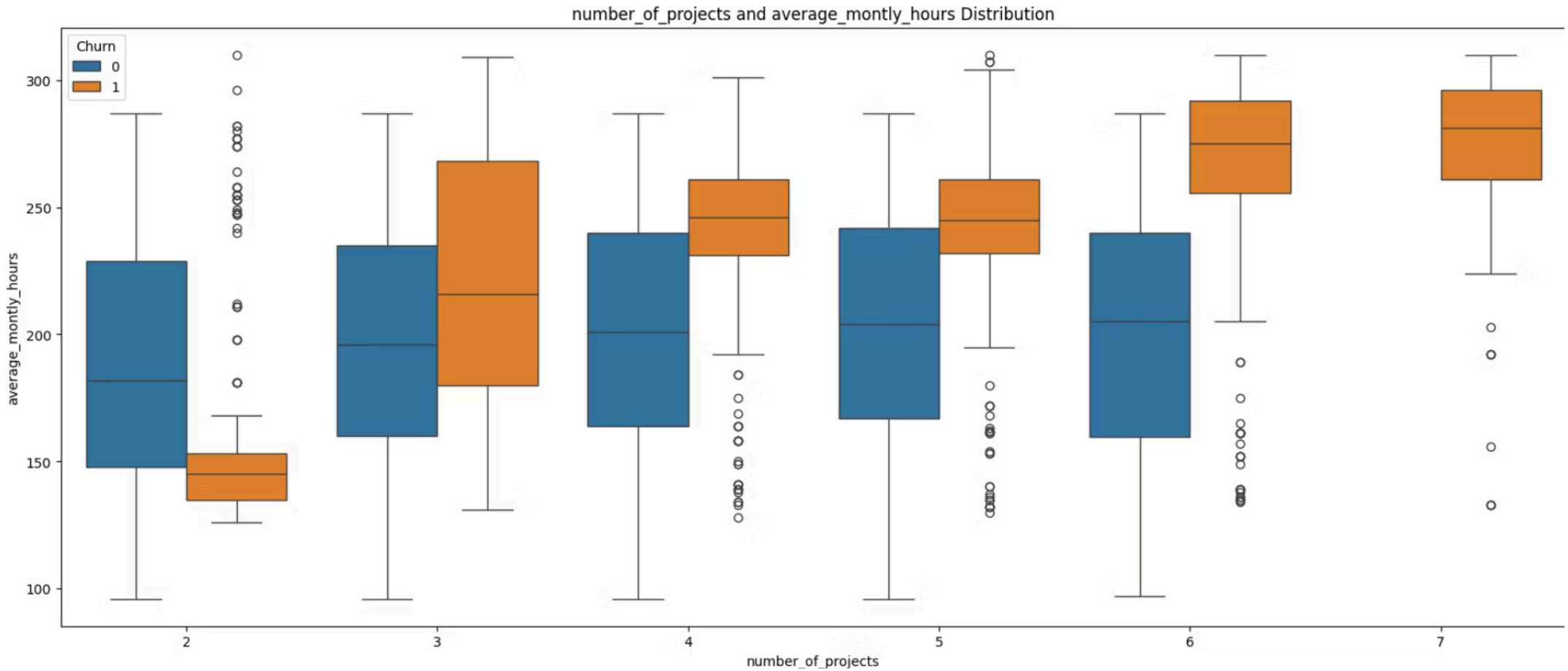
Department, Salary_INR, and Churn



Departments with the highest churn, such as Sales and Support, tend to offer lower salaries. Meanwhile, departments like Management and R&D offer better pay and see less churn.

Key Churn Indicators

number_of_projects and average_montly_hours and Churn



Employees with **very high project counts (6–7)** also record **longer average monthly hours**. This combination significantly increases **churn rate**, likely due to overwork and stress.

On the other hand, employees with **3–5 projects** and **moderate hours** show the **lowest churn**.

Data Preprocessing

Preparing the data for model training involves several key steps:

Feature and Target Definition

Separated employee attributes (X) from the churn target variable (y).

Data Encoding

Converted categorical 'Department' data into numerical format using Label Encoding.

Data Scaling

Normalized 'average_montly_hours' and 'Salary_INR' to a common scale using MinMaxScaler.

Data Splitting

Divided the dataset into training (80%) and testing (20%) sets for model evaluation.

Model Performance Benchmarks

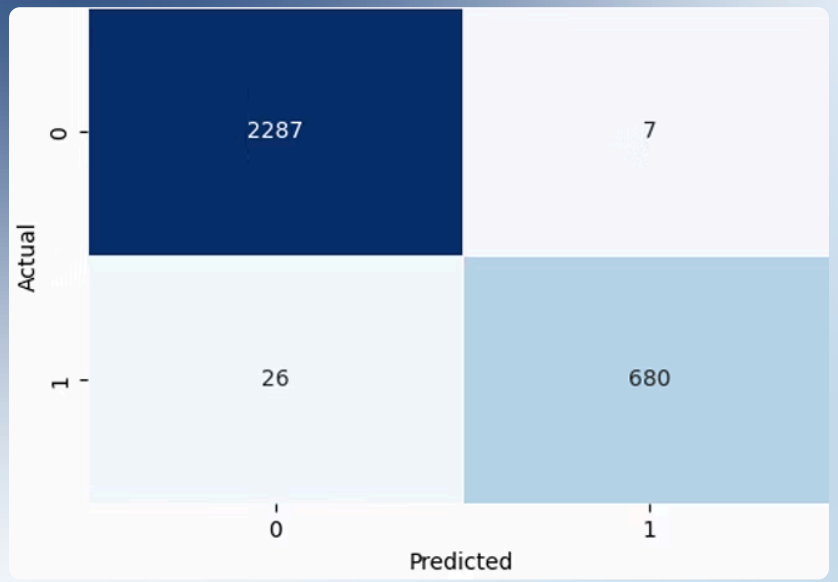
This table summarizes the performance of various machine learning models, from foundational algorithms to advanced ensemble and boosting methods, in predicting employee churn.

98.87%	98.73%	98.53%	98.53%
StckingClassifier	ExtraTreesClassifier	RandomForestClassifier	LGBMClassifier
98.50%	98.23%	98.17%	97.50%
XGBoostClassifier	CatBoostClassifier	BaggingClassifier	Decision Tree
95.83%	94.63%	77.13%	76.40%
KNeighborsClassifier	SVC	Logistic Regression	LinearSVC

The evaluation demonstrates that ensemble and boosting methods, particularly ExtraTreesClassifier, VotingClassifier, RandomForestClassifier, LGBMClassifier, and XGBoostClassifier, achieve the highest accuracy in predicting employee churn. Foundational models like Logistic Regression and LinearSVC show significantly lower performance, highlighting the benefit of more sophisticated algorithms for this prediction task.

VotingClassifier Accuracy

98.90%



Classification Report: Model Performance

The classification report details the precision, recall, F1-score, and support for each class (0 for non-churn, 1 for churn).

The model demonstrates high performance, with an overall accuracy of 99%. This indicates a strong ability to correctly predict both employees who will churn and those who will not.

Class	Precision	Recall	F1-Score	Support
0 (No Churn)	0.99	1.00	0.99	2294
1 (Churn)	0.99	0.96	0.98	706
Total	0.99	0.98	0.98	3000