Data Analysis Final Project

IBM HR Analytics Employee Attrition & Performance

Fighting Illini 🌽

Zihan Li, Yajing Gao, Yuelin Zou, Yiting Wang, Tongxin Liu

Agenda

- Background
- Data Preprocessing
- Further Improvement
- Data Analysis
- Insight
- Application & Limitation

Background

Background

Proper labor turnover rate can promote their own development

An **overly** high labor turnover rate will bring many adverse effects

Utilize **right** methods to solve the issue of how to retain employees

Word Clouds Preview

Pros

Cons

Data Preprocessing

View of Data

- Dependent Variable: Attrition (Yes/No)

  Factors that influence attrition in a company

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical

Data Cleaning and Preprocessing

1. Explore data:
   a. No missing values
   b. 1470 rows/ 35 columns
   c. 26 numerical, 9 categorical variables
2. Convert our target variable - " attrition" column to the first column
3. Drop all the columns that we do not need :
   - Values are all the same (columns with only one unique values)
   - Columns has no analyzing values (employee id)
   - Values that has high correlation (high similarities)
   - 35 columns → 28 columns

```
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64
 9   EmployeeNumber            1470 non-null   int64
 10  EnvironmentSatisfaction   1470 non-null   int64
```

EmployeeCount	EmployeeNumber
1	1
1	2
1	4
1	5
1	7

Over18
Y
Y
Y
Y
Y

StandardHours
80
80
80
80
80

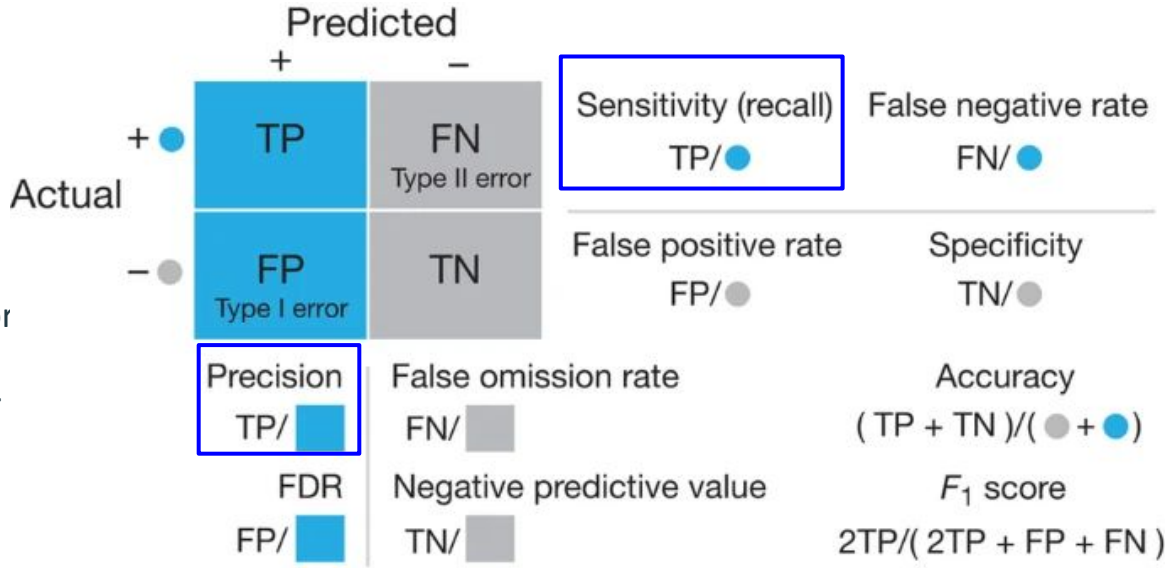
```
['DailyRate','HourlyRate','MonthlyRate']
```

Further Improvement

Indicator Selection

Selected indicators:

1. **Recall Rate (Sensitivity):**
   - How well our model is able to identify the unstable employees from the company
1. **Precision:**
   - supplementary indicator
1. **AUC:**
   - the ability of a classifier to distinguish between classes
1. **F-score:**
   - captures both the recall and precision results

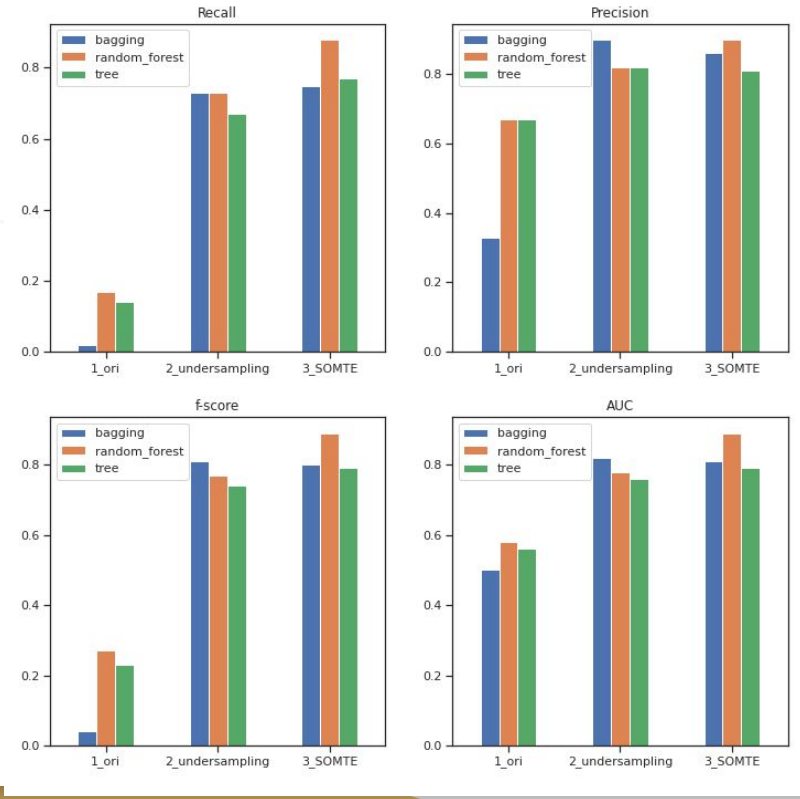


Imbalanced dataset adjustment

To solve the unbalanced data problem:

1. Near Miss Undersampling
2. SMOTE: Synthetic Minority Over-sampling Technique

		Recall	Precision	f-score	AUC
Iteration_1_original	tree	0.14	0.67	0.23	0.56
	random_forest	0.17	0.67	0.27	0.58
	bagging	0.02	0.33	0.04	0.50
Iteration_2_undersampling	tree	0.67	0.82	0.74	0.76
	random_forest	0.73	0.82	0.77	0.78
	bagging	0.73	0.90	0.81	0.82
Iteration_3_SMOTE	tree	0.77	0.81	0.79	0.79
	random_forest	0.88	0.90	0.89	0.89
	bagging	0.75	0.86	0.80	0.81



One Hot Encoding

- Eliminating inherent ordering

	BusinessTravel	Department	EducationField	Gender	JobRole	MaritalStatus	OverTime
0	Travel_Rarely	Sales	Life Sciences	Female	Sales Executive	Single	Yes
1	Travel_Frequently	Research & Development	Life Sciences	Male	Research Scientist	Married	No
2	Travel_Rarely	Research & Development	Other	Male	Laboratory Technician	Single	Yes
3	Travel_Frequently	Research & Development	Life Sciences	Female	Research Scientist	Married	Yes
4	Travel_Rarely	Research & Development	Medical	Male	Laboratory Technician	Married	No

Direct mapping

Business Travel	
Non-Travel	1
Travel_Rarely	2
Travel_Frequently	3

	BusinessTravel
0	3
1	2
2	3
3	2
4	3

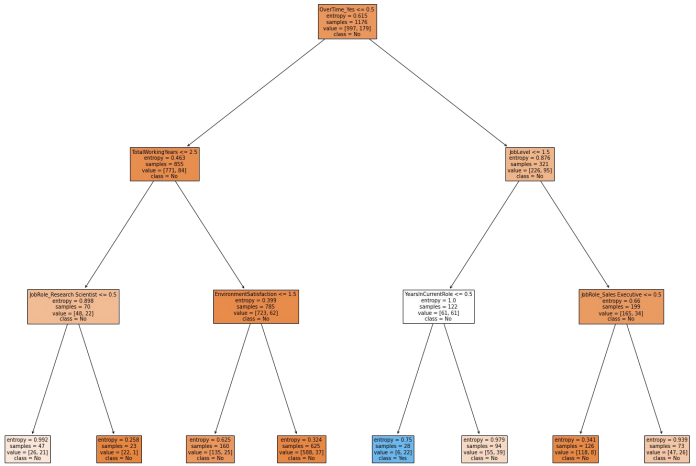
One hot encoding

	BusinessTravel_Travel_Frequently	BusinessTravel_Travel_Rarely
0	0	1
1	1	0
2	0	1
3	1	0
4	0	1

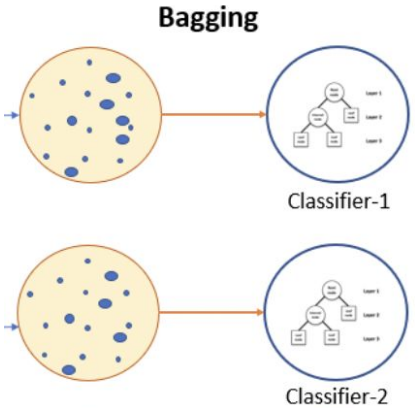
Data Analysis With Machine Learning

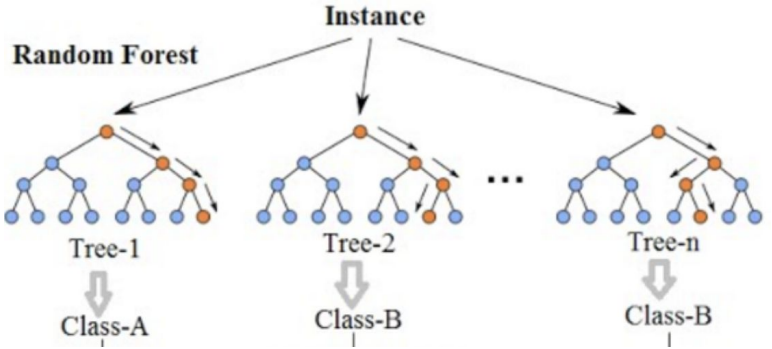
Model Selection

Decision Tree

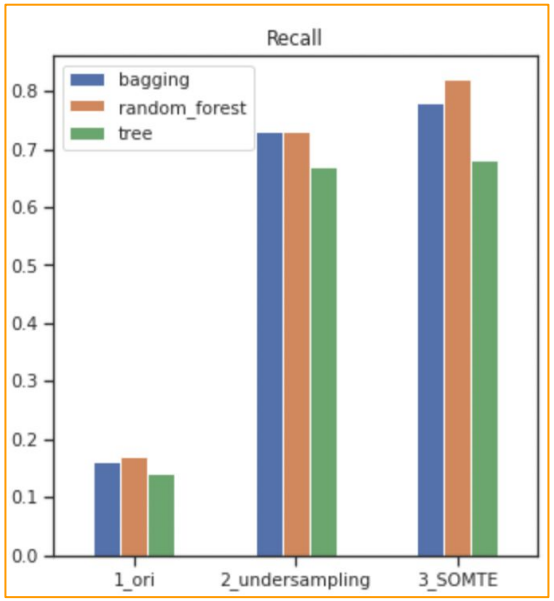


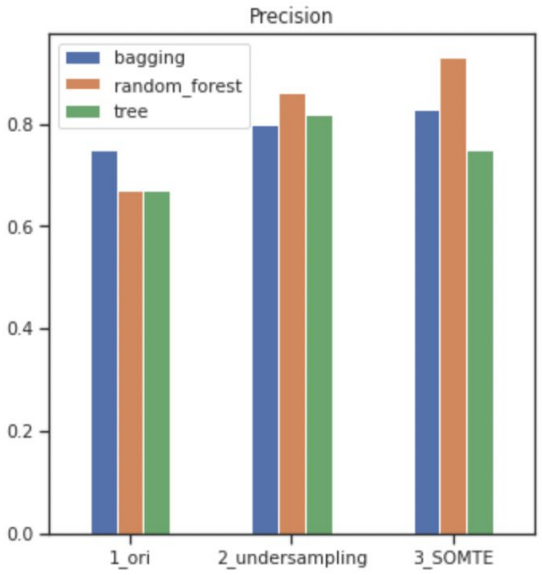
- Easy to use
- Explanatory

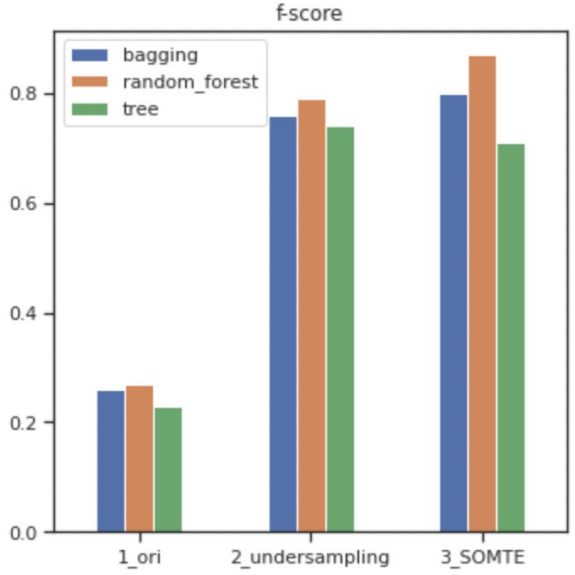




Iteration Comparison







Handling Imbalanced dataset greatly improve the performance of the models

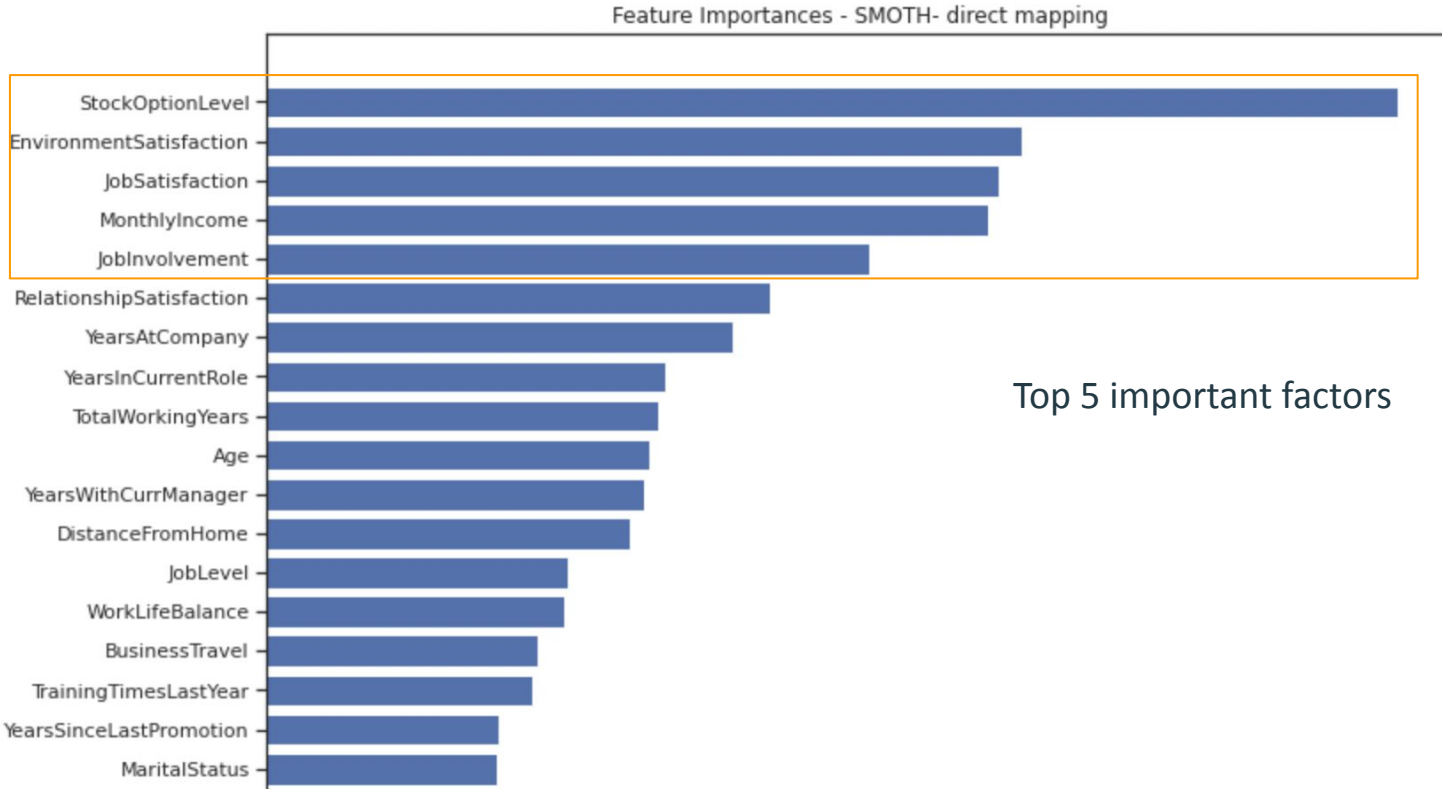
One Hot Encoding

		Recall	Precision	f-score	AUC
Iteration_1_original	tree	0.14	0.67	0.23	0.56
	random_forest	0.17	0.67	0.27	0.58
	bagging	0.16	0.75	0.26	0.57
Iteration_2_undersampling	tree	0.67	0.82	0.74	0.76
	random_forest	0.73	0.86	0.79	0.80
	bagging	0.73	0.80	0.76	0.77
Iteration_3_SMOTE	tree	0.68	0.75	0.71	0.72
	random_forest	0.82	0.93	0.87	0.88
	bagging	0.78	0.83	0.80	0.81
Iteration_4_one_hot	random_forest	0.83	0.96	0.89	0.90

Do have small Improvement

Insight

Top Attributes - Random Forest(SMOTH)

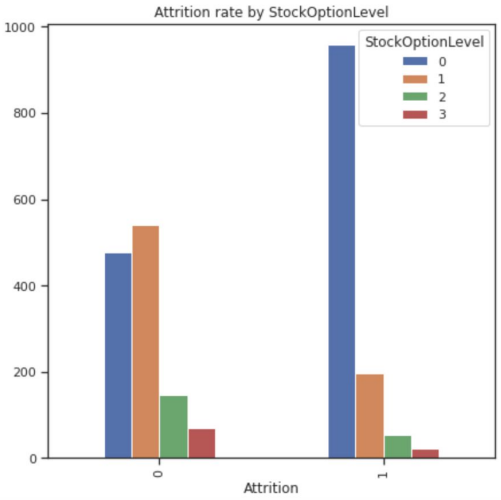


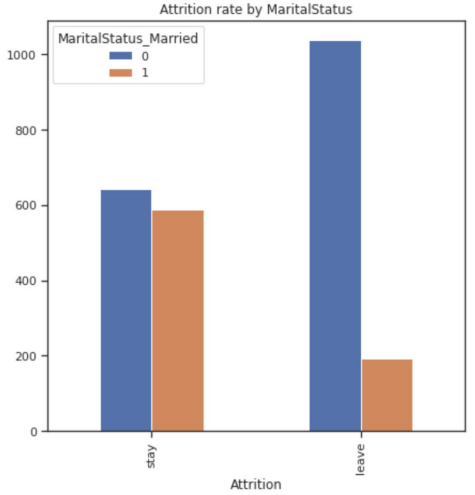
Top 5 important factors

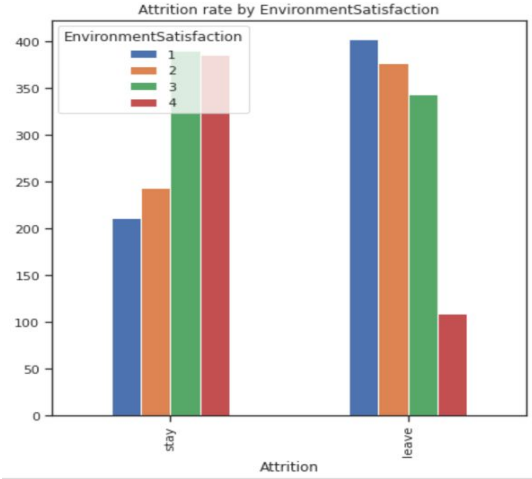
Insights

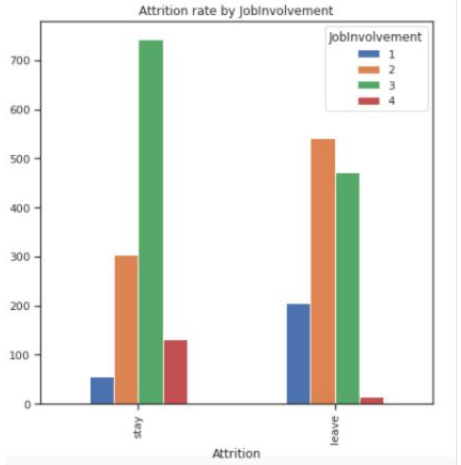
What kind of person are more likely to leave the company?

Unmarried employees with **low stock options** who **don't enjoy with their working environments** and are **not engaged in their jobs** tend to leave the company



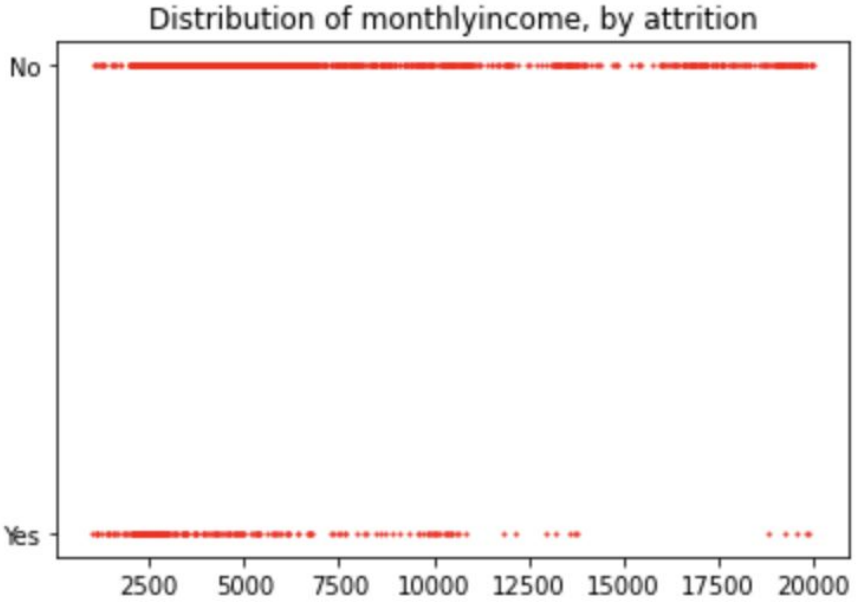


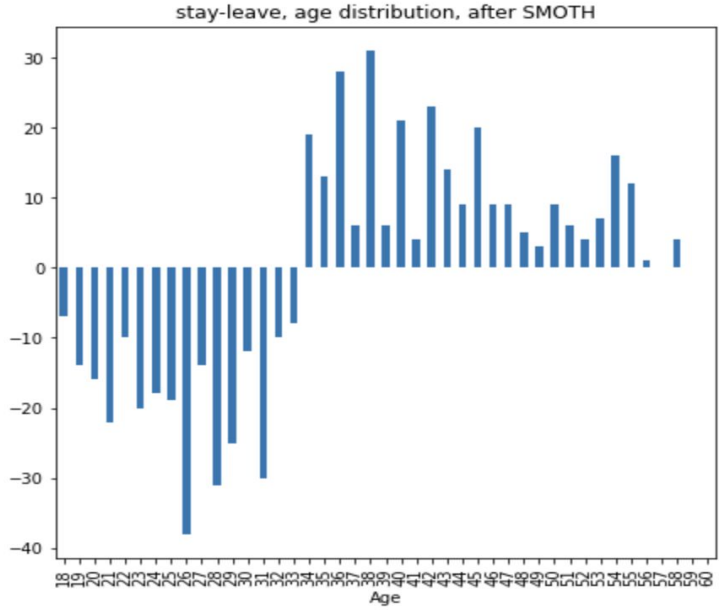




Insights

Age and monthly income are also key factors determining the attrition rate of employees





Employees **under 33 years of age** with **low monthly income** tend to leave the company

Application and Limitation

Application & Limitations

01 **Analyze Attrition Rate**
hiring process, outsourcing personnels

02 **Important Factors**
work-life balance, monthly salary, years since last promotion

03 **Improve These Factors**
job and working environment satisfaction

04 **Better Understanding**
what employees want, improve cohesion

01 **Sensitive Topic**
not for promotion or hiring decision

02 **Limited Application**
differences between companies

03 **Biased Sampling**
more attrition than not attrition

04 **Limited Data**
reduce accuracy of model

Thanks for your listening