SAMSUNG

Together for Tomorrow! Enabling People

Education for Future Generations

Samsung Innovation Campus

Artificial Intelligence Course

Employee Future.

Project presented by :

- Abdallah Hassan.
- Abdelrahman Morsi.
- Ahmed Alaa.

- Facilitator Mohamed Abdallah

Team : Triple A
Data Used : Employee Future

Table of content :

About our project :

Our project is about evaluating and building a predictive model for a dataset for a company's HR department to fulfill their objective which we will later on state.

About the Data:

Predict the employee's future in the company based on below data:

- 1- **Education**: The education level (Bachelors, Master, PHD).
- 2- **JoiningYear**: The year of joining the company between 2012 and 2018.
- 3- **City**: City office where posted (Bangalore, Pune, New Delhi).
- 4- **PaymentTier**: Payment is divided into 3 levels.

About the Data:

- 5- **Age**: The age of the employee (between 22 and 41).
- 6- **Gender**: male or female.
- 7- **EverBenched**: Ever stay out of projects for a month or more.
- 8- **ExperienceInCurrentDomain:** the Experience of the employee In Current Domain.

Objective:

Our job is to build a predictive model that predicts the prospects of future and present employee.
In other words, we want to predict whether the employee will leave or not in next 2 years.

EDA:

First, we need to take a look at our data information :

	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	ExperienceInCurrentDomain	LeaveOrNot
0	Bachelors	2017	Bangalore	3	34	Male	No	0	0
1	Bachelors	2013	Pune	1	28	Female	No	3	1
2	Bachelors	2014	New Delhi	3	38	Female	No	2	0
3	Masters	2016	Bangalore	3	27	Male	No	5	1
4	Masters	2017	Pune	3	24	Male	Yes	2	1

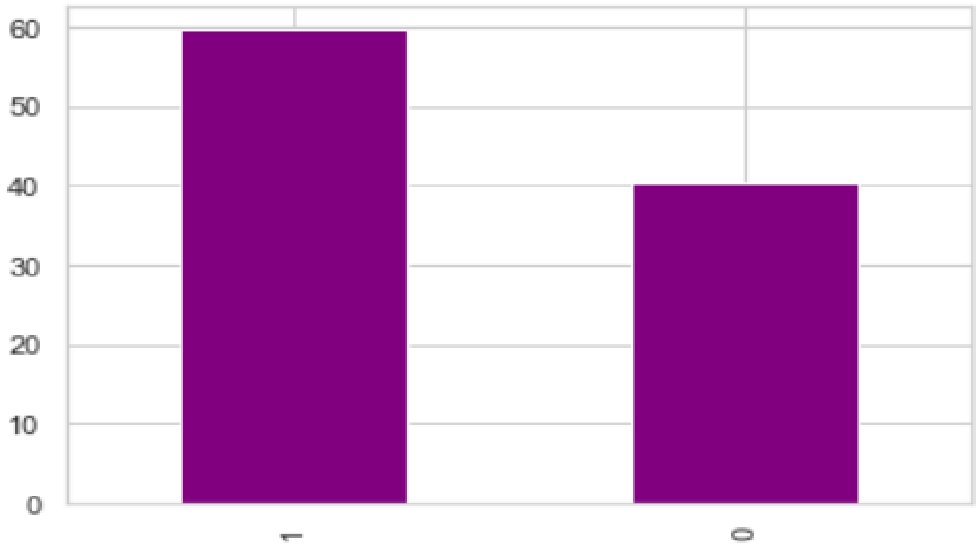
EDA:

We also need to take a look at the description of the data :

	JoiningYear	PaymentTier	Age	ExperienceInCurrentDomain	LeaveOrNot
count	4653.000000	4653.000000	4653.000000	4653.000000	4653.000000
mean	2015.062970	2.698259	29.393295	2.905652	0.343864
std	1.863377	0.561435	4.826087	1.558240	0.475047
min	2012.000000	1.000000	22.000000	0.000000	0.000000
25%	2013.000000	3.000000	26.000000	2.000000	0.000000
50%	2015.000000	3.000000	28.000000	3.000000	0.000000
75%	2017.000000	3.000000	32.000000	4.000000	1.000000
max	2018.000000	3.000000	41.000000	7.000000	1.000000

EDA:

Let's take a look at the ratio between Male and Female employees in our company.



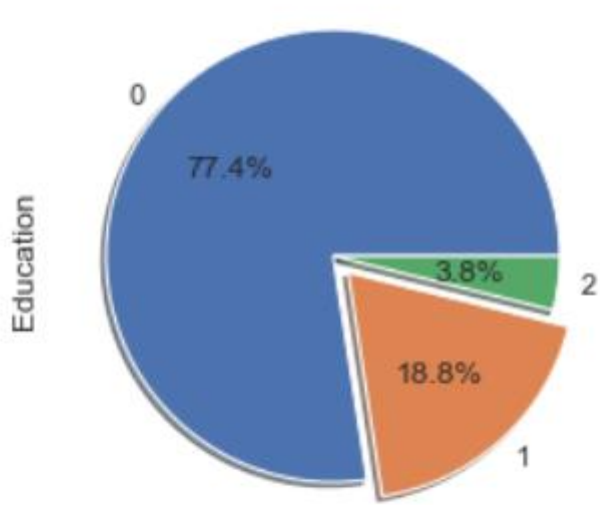
1: Male 0:Female

As we can see here the number of males in the company is slightly higher than females.

EDA:

Here we also used a pie chart to see the ratio between education levels in our company.

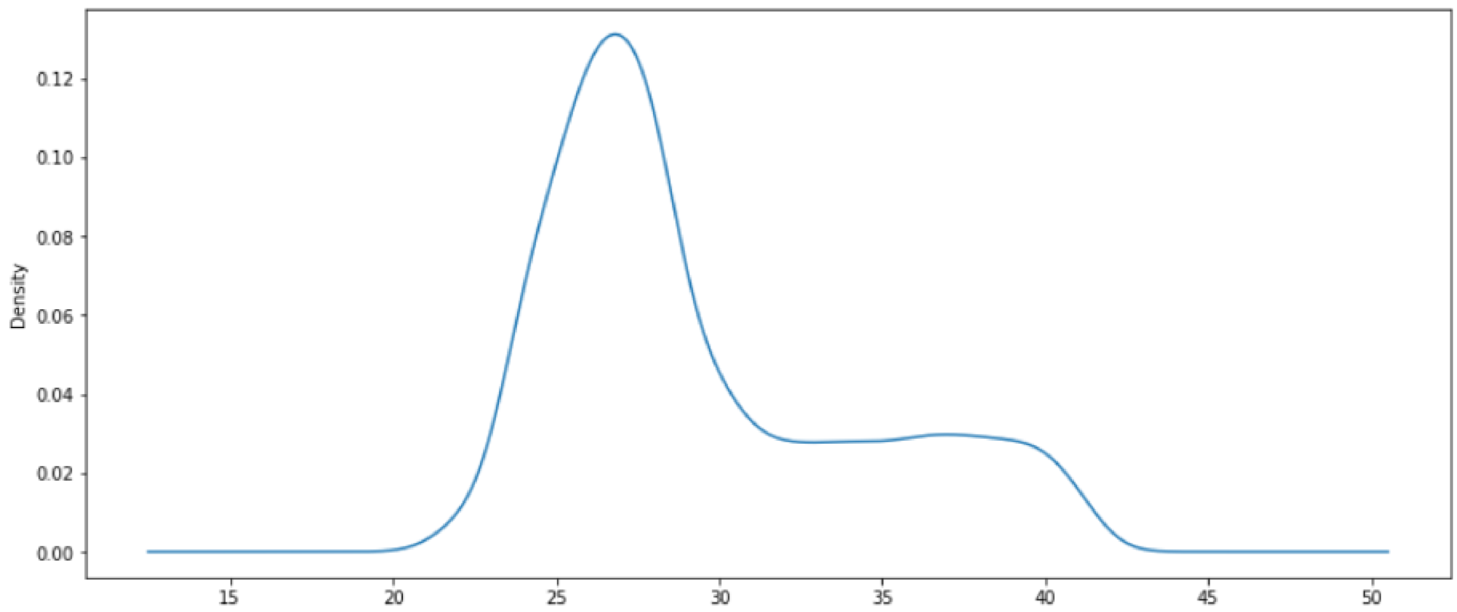
0: Bachelors
1: Masters
2: PHD



It is expected to have a low percentage of PHDs as it takes 3-7 years to complete

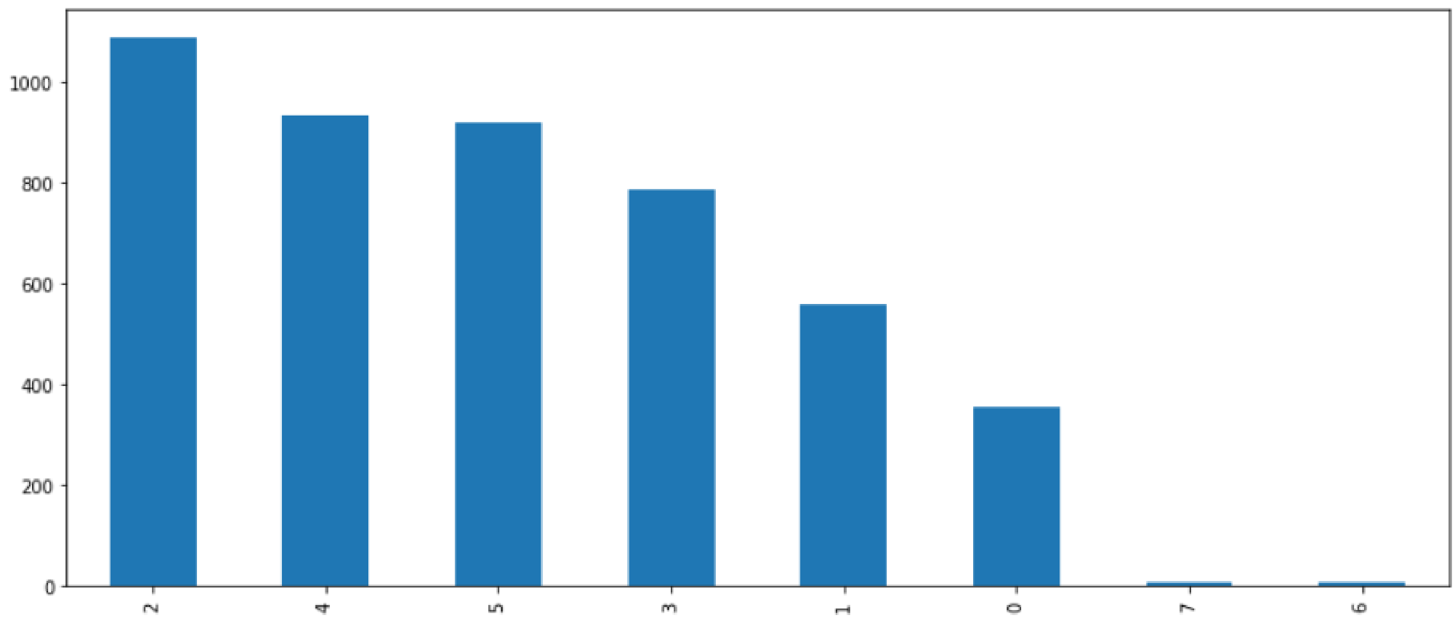
EDA:

The distribution of employee ages in our company.



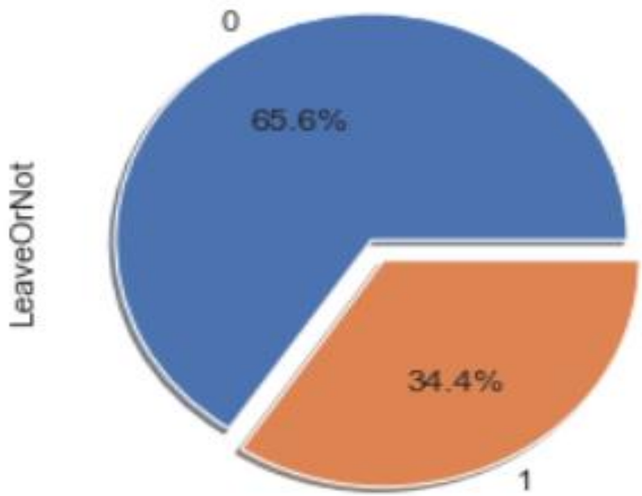
EDA:

The number of employee due to experience in current domain.



EDA:

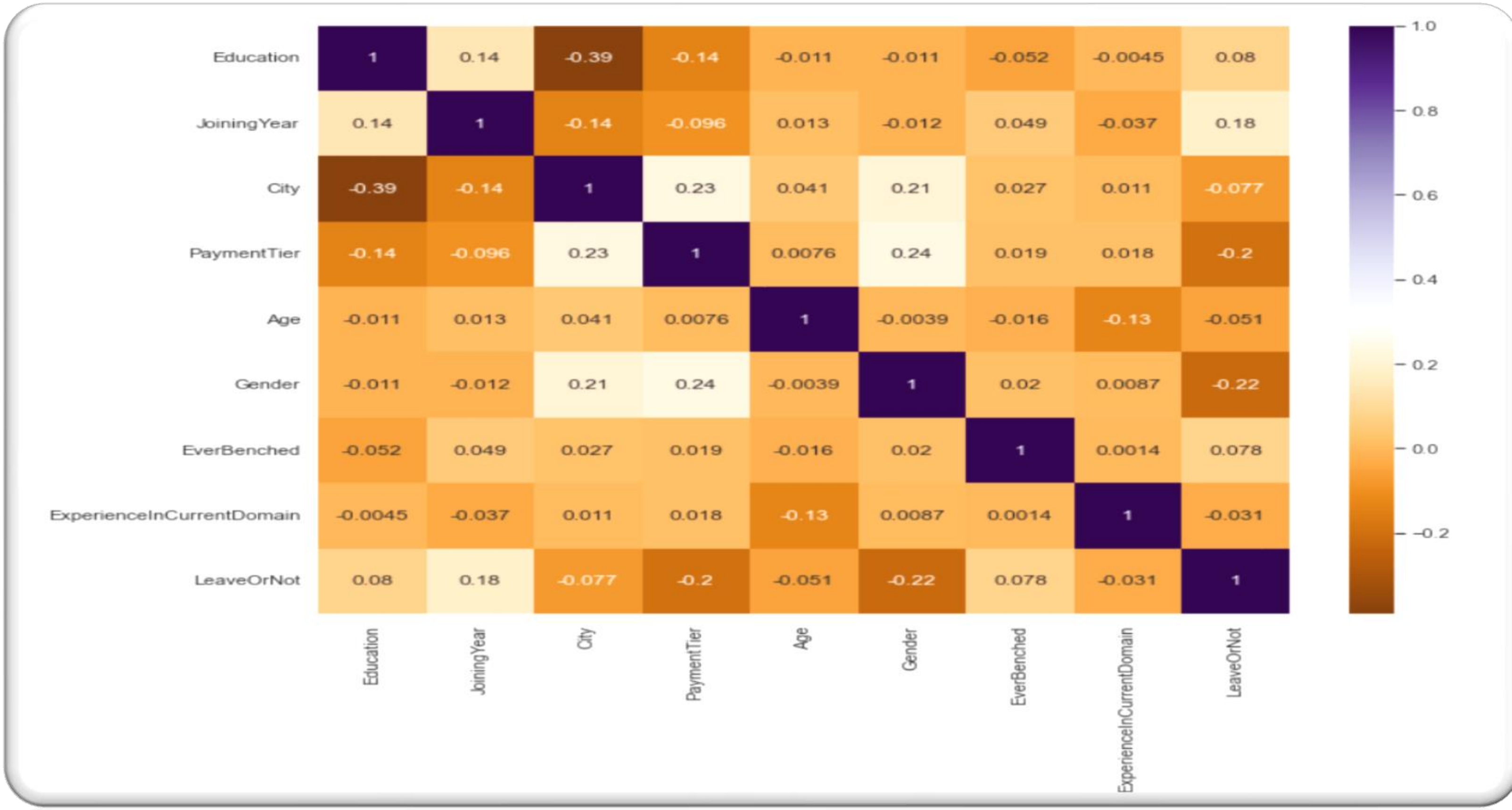
A pie chart to see the ratio between employees who left and those who stayed



0: Stayed 1: Left

EDA:

From this correlation heatmap we conclude that there is no strong relations between columns.



Preprocessing:

=> Dealing with Duplicates :-

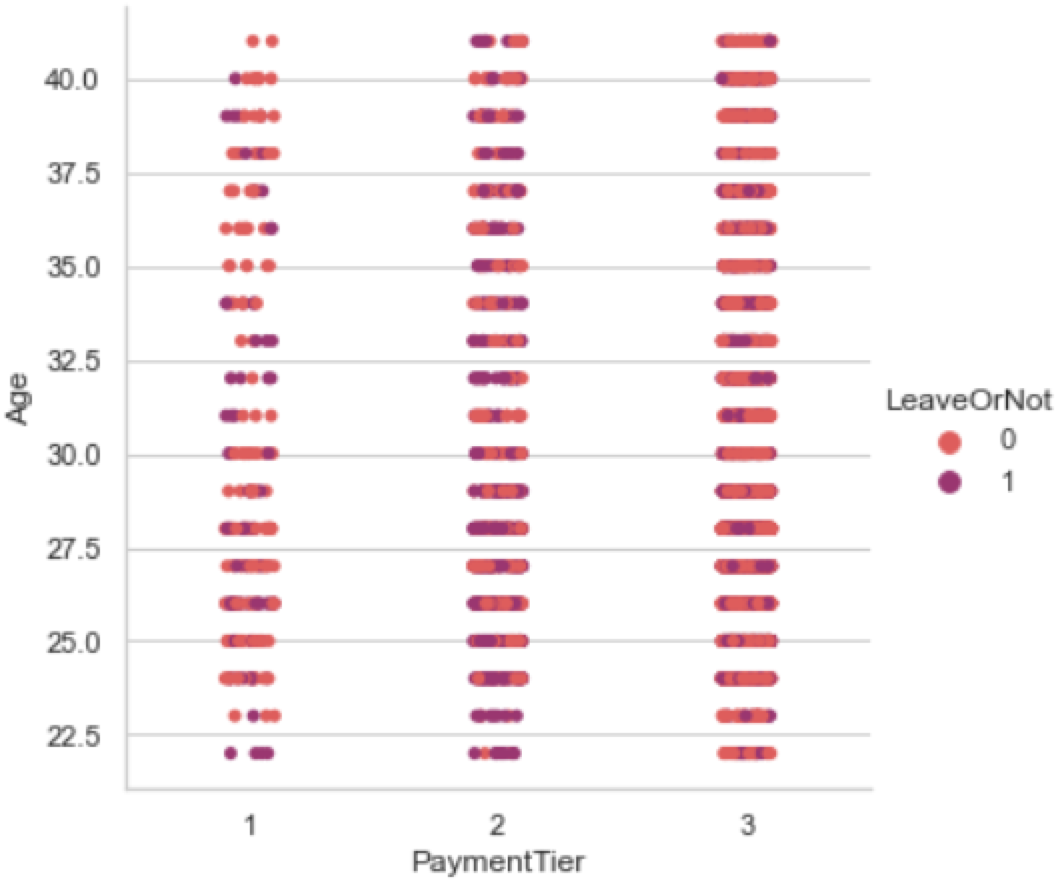
After inspecting the ratio between the dropped data and the whole data we found out that we would have to drop 40% of our data which will later on affect our model accuracy. So we will not drop the duplicates.

=> Dealing with Missing data :-

No missing data.

```
Education                    0
JoiningYear                  0
City                         0
PaymentTier                  0
Age                          0
Gender                       0
EverBenched                  0
ExperienceInCurrentDomain    0
LeaveOrNot                   0
dtype: int64
```

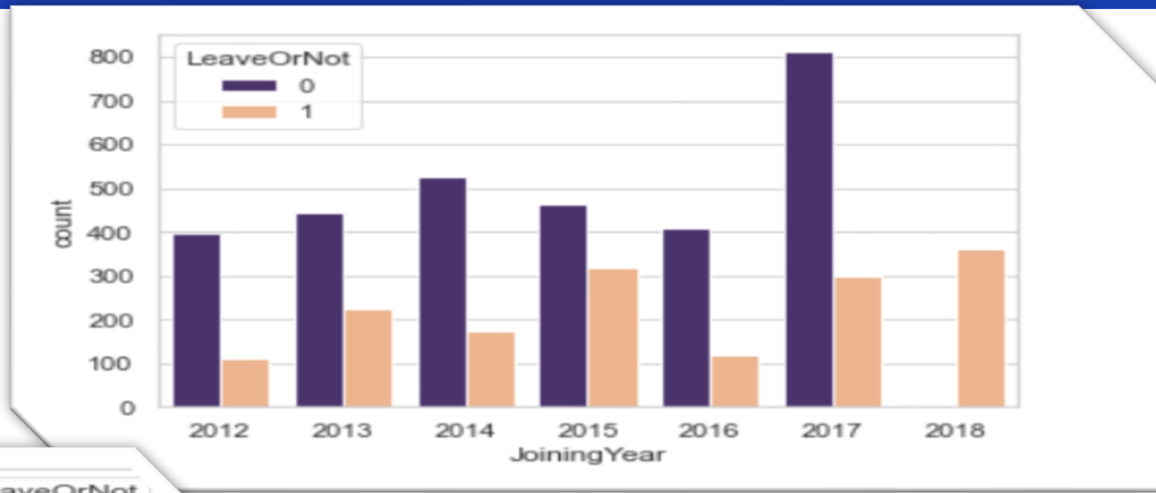
Business Solution:



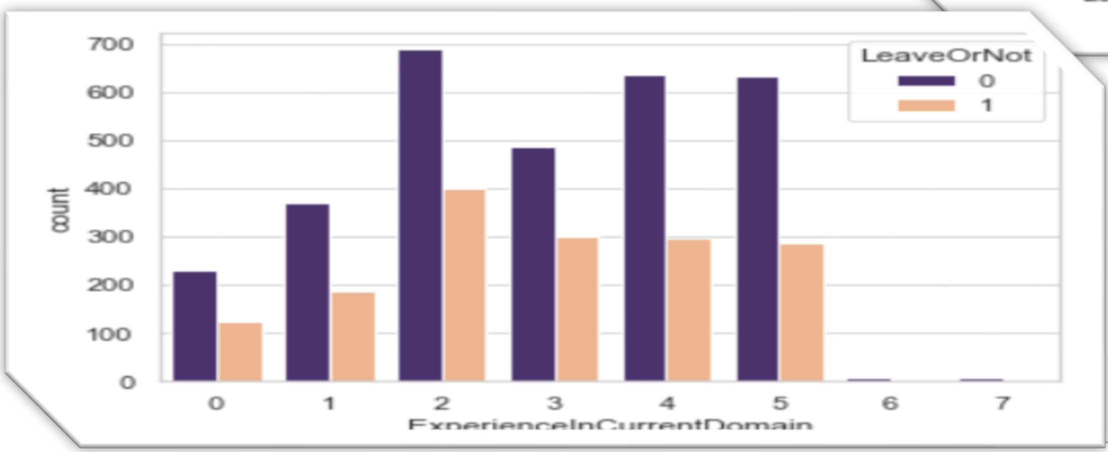
Here we are plotting the relationship between payment tiers and experience in current domain with respect to leaving or staying in the company using scatter plot.

Observation: More people at age from 22.5 to 26 leave the company in payment tier 2 than any other tier.

Business Solution:

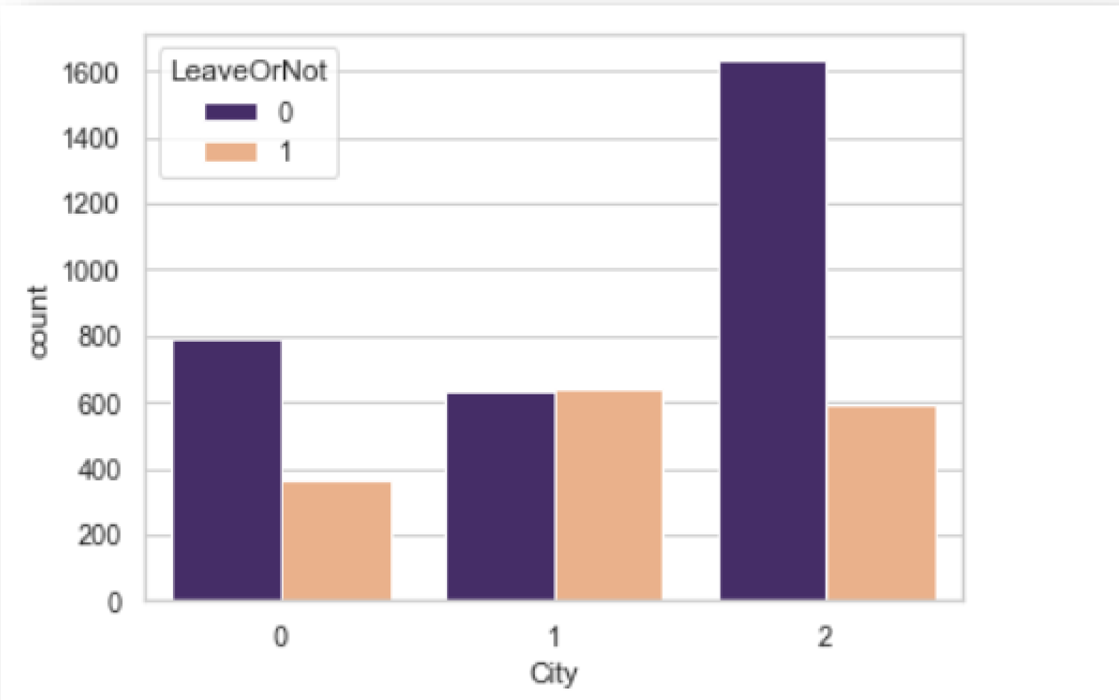


In 2018 most of the employees left which is a bad sign so, we need to review what went wrong this year in particular.(CRITICAL PROBLEM!!)



- Employees with 2 years of experience in their domain are the most to leave.

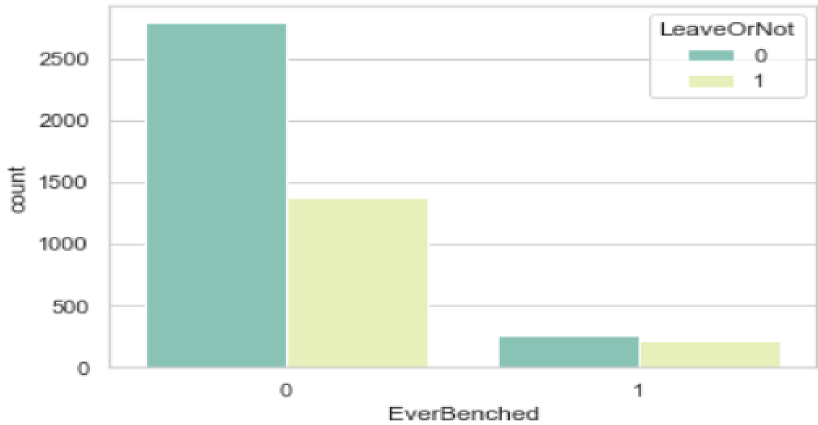
Business Solution:



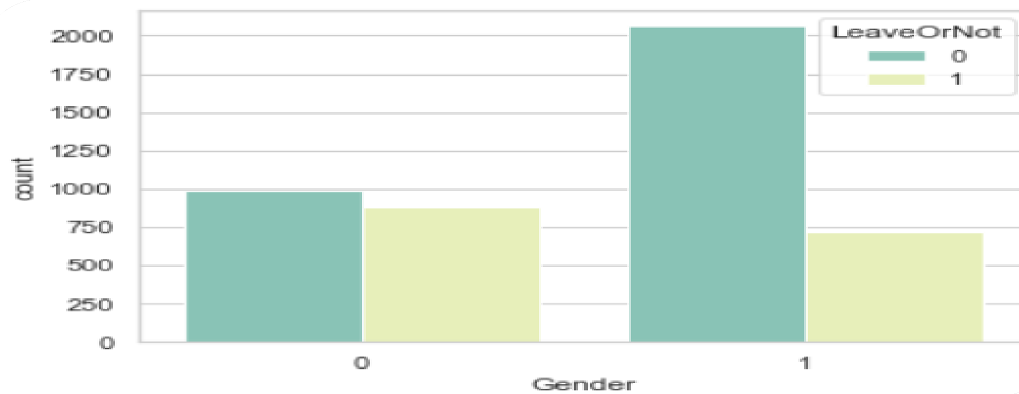
Most employees that live in Pune leave (more than 50% of employees living there) and it is also the most city that employees leave our company from followed by Bangalore then New Delhi.

‘Bangalore’ : 2 ‘Pune’ : 1 ‘New Delhi’ : 0

Business Solution:



Most Employees who get benched leave (approximately 45%)
0: Not benched   1: Benched



- A lot of females tend to leave our company.

Business Solution:

So, after asking business questions and answering them to filter out the problems within our company's dataset.

How can we solve them?

Business Solution Conclusion:

Pros and cons:

Pros:
•Not many employees get benched.
•23% of our employees have a masters degree.
•Not many employees from New Delhi leave our company.

Cons:

- A lot of females tend to leave our company.
- Most employees that live in Pune leave (more than 50% of employees living there)
- In 2018 most of the employees left.
- Employees with 2 years of experience in their domain are the most to leave.

Business Solution Conclusion:

Solving company's problems:

•Make the work environment more suitable for females to work in.
•Supply employees from Pune and Bangalore more means of transportation.
•Try encourage employees aged from 22 till 26 to stay by rewarding them with bonuses or ranking them up in Payment tiers.
•Review what went wrong in 2018 because nearly all employees left the company. (CRITICAL PROBLEM)
•Try to lower down employee benching rate by rotating more employees in projects.
•Encourage employees with 2 years of experience of stay by involving them in more projects to gain more experience and also get paid more.

Modeling:

- We split the data based on (the features & the target , train & test) and then we started the algorithm tests to choose from.

- **The classification algorithm we used is:**

1. KNeighborsRegressor
2. DecisionTreeRegressor
3. XGBRegressor
4. SupportVectorMachine

- We found that the best models are KNN and SVM so we will take you on a tour to find out their results.

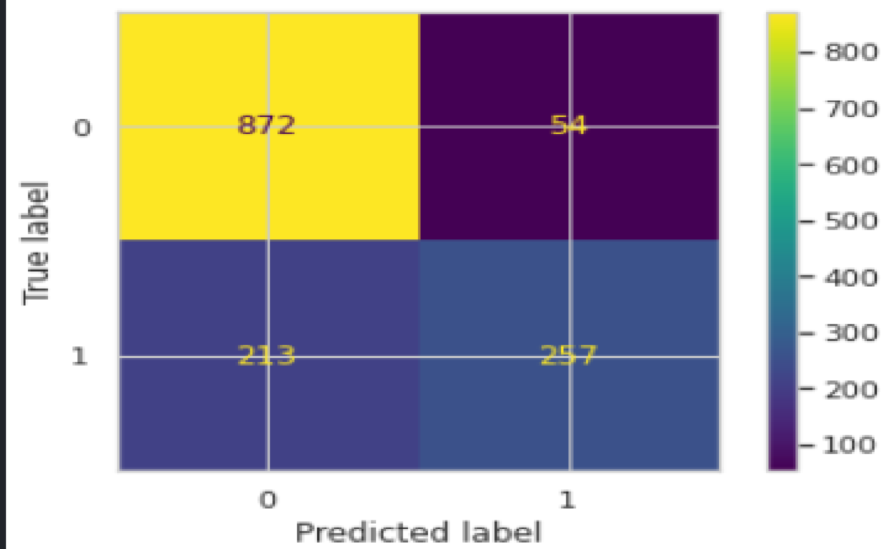
Modeling:

KNN

Modeling:

KNN

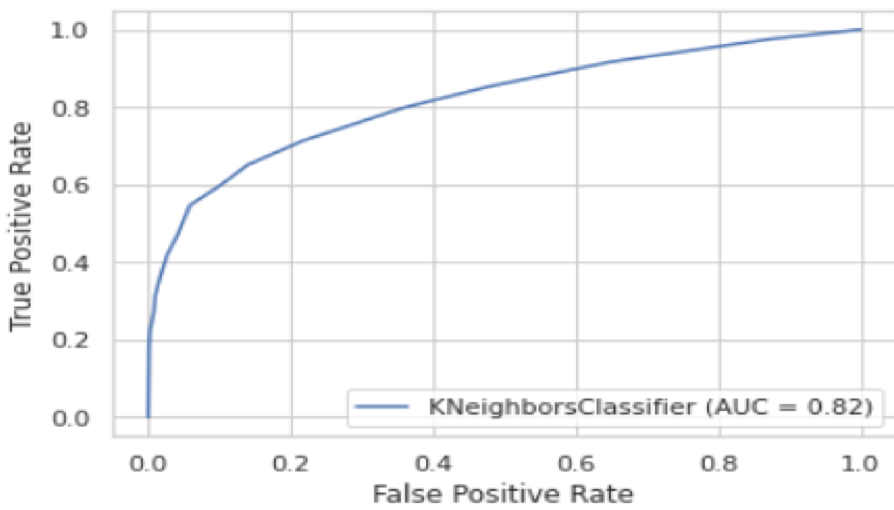
```
/n
              precision    recall  f1-score   support

           0       0.80      0.94      0.87       926
           1       0.83      0.55      0.66       470

    accuracy                           0.81      1396
   macro avg       0.82      0.74      0.76      1396
weighted avg       0.81      0.81      0.80      1396

/n
```

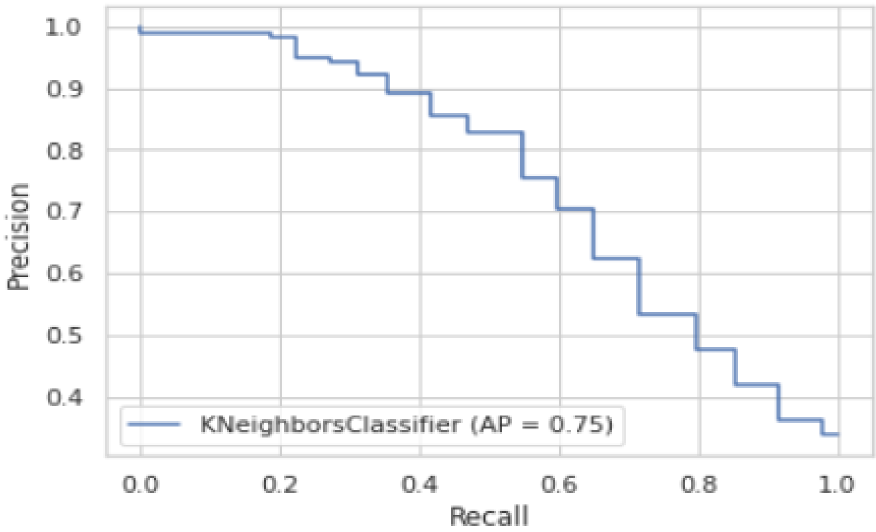


Here we completed our classification model by showing the classification report and the confusion matrix to review our model.

Modeling:

KNN



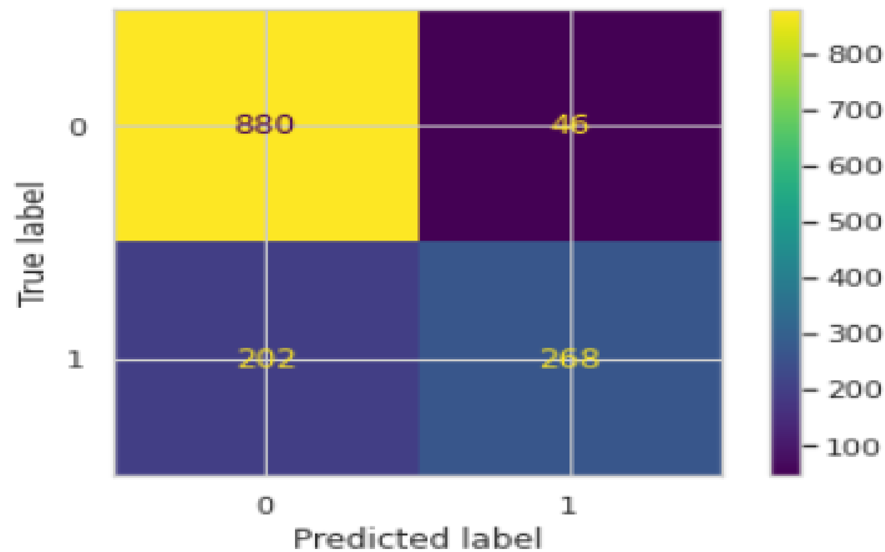


We are previewing the ROC curve and the Precision-Recall curve to know the exactness of our model.

Modeling:

SVM

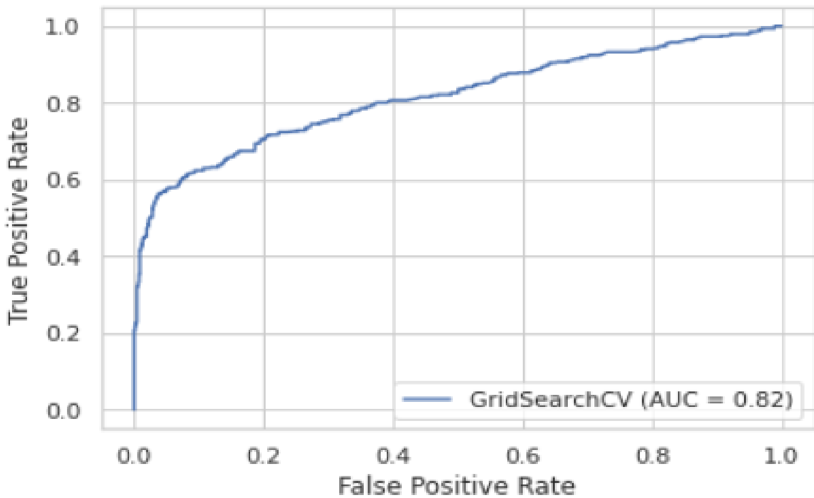
```
/n
              precision    recall  f1-score   support

           0       0.81      0.95      0.88       926
           1       0.85      0.57      0.68       470

    accuracy                           0.82      1396
   macro avg       0.83      0.76      0.78      1396
weighted avg       0.83      0.82      0.81      1396

/n
```

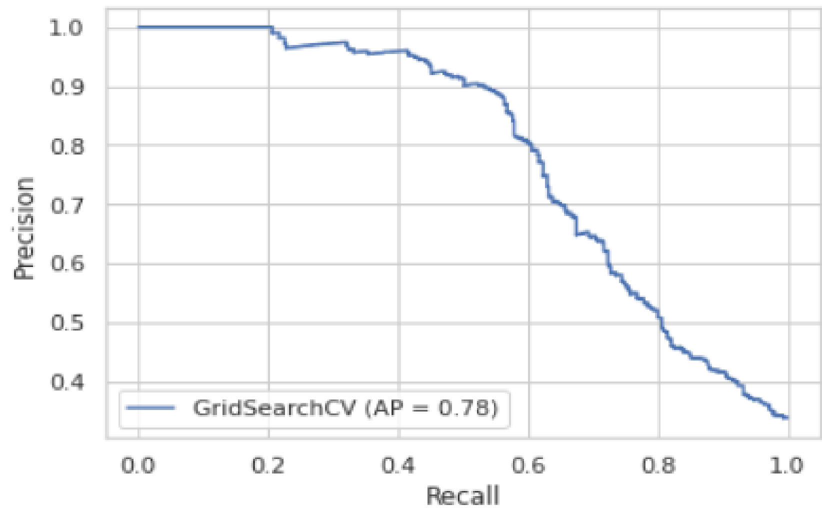


Like the previous one, we showing the classification report and the confusion matrix to review our model.

Modeling:

SVM





We are previewing the ROC curve and the Precision-Recall curve to know the exactness of our model.

Modeling Conclusion:

- To sum it all up, using SVM and KNN was the best choice between several other modeling algorithms as mentioned before.
- Accuracy of KNN was: 81%
- Accuracy of SVM was: 82%
- And this was the best accuracy that we got from this dataset

Links:

- [GitHub Link]
- [Kaggle Link]
- [Presentation]

Questions?

Thank you

SAMSUNG

Together for Tomorrow!

Enabling People

Education for Future Generations