PROJECT REPORT

Attrition rate in the company

Submitted towards the partial fulfillment of the criteria for award of Genpact Data Science Prodegree by Imarticus

Submitted By:
Sumit Daniel

Course and Batch:
Data Science Pro degree - 31

\

Acknowledgements

We are using this opportunity to express my gratitude to everyone who supported us throughout the course of this group project. We are thankful for their aspiring guidance, invaluably constructive criticism, and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on several issues related to the project.

Further, we were fortunate to have **Dr. Vinod Murti** as our mentor. He has readily shared his immense knowledge in data analytics and guide us in a manner that the outcome resulted in enhancing our data skills.

We wish to thank, all the faculties, as this project utilized knowledge gained from every course that formed the DSP program.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: April 01, 2018 Sumit John Daniel

Place: Mumbai

Certificate of Completion

I hereby certify that the project titled **"Attrition rate in the company"** was undertaken and completed under my supervision by **Sumit Daniel** from the batch of DSP 31.

Mentor: Dr. Vinod Murti

Date: 18 December, 2020

Place – Mumbai

CHAPTER 1: INTRODUCTION

Introduction of the topic

Attrition rate -:

A common attrition rate definition refers to employee or staff turnover, but in a broader sense, attrition rate is a calculation of the number of individuals or items that vacate or move out of a larger, collective group over a specified time frame.

Attrition rate is also commonly referred to as churn rate. A term often used by human resources professionals to determine a company's ability to retain employees, attrition rate is increasingly used in the marketing world as a figure that points to the company's ability to retain customers or to project the number of new sales necessary to maintain the status quo, accounting for customer churn or customer attrition.

Major Reasons for Attrition

- Relationship with the boss
- Bored and unchallenged by the work itself
- Relationships with coworkers
- Opportunities to use their skills and abilities
- Contribution of their work to the organization's business goals
- Autonomy and independence on the job
- Meaningfulness of the employee's job
- Knowledge about your organization's financial stability
- Overall corporate culture
- Management's recognition of employee job performance

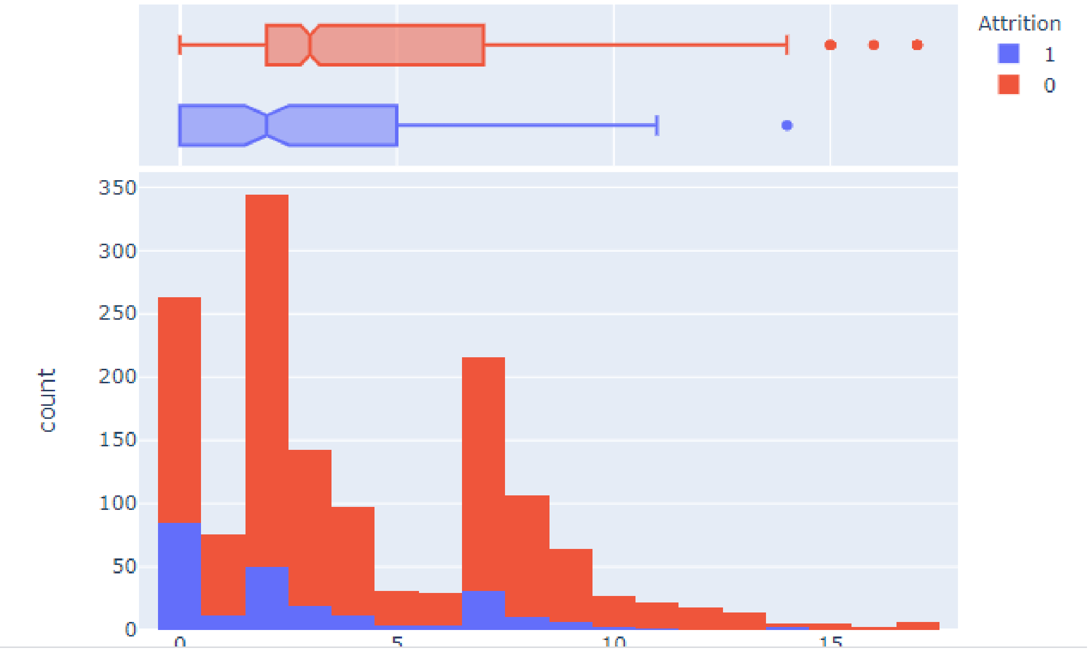
Objective of the study

- To understand the major reason for attrition of employee in the company
- To get the accuracy of the attrition rate
- To build the models with logistic regression, Random forest and SMOTE

Exploratory Data Analysis

The employee does not necessarily have to establish a friendly relationship with the boss, but it is necessary that good communication exists. According to many sources: A bad boss is also the number one reason why employees quit their job.

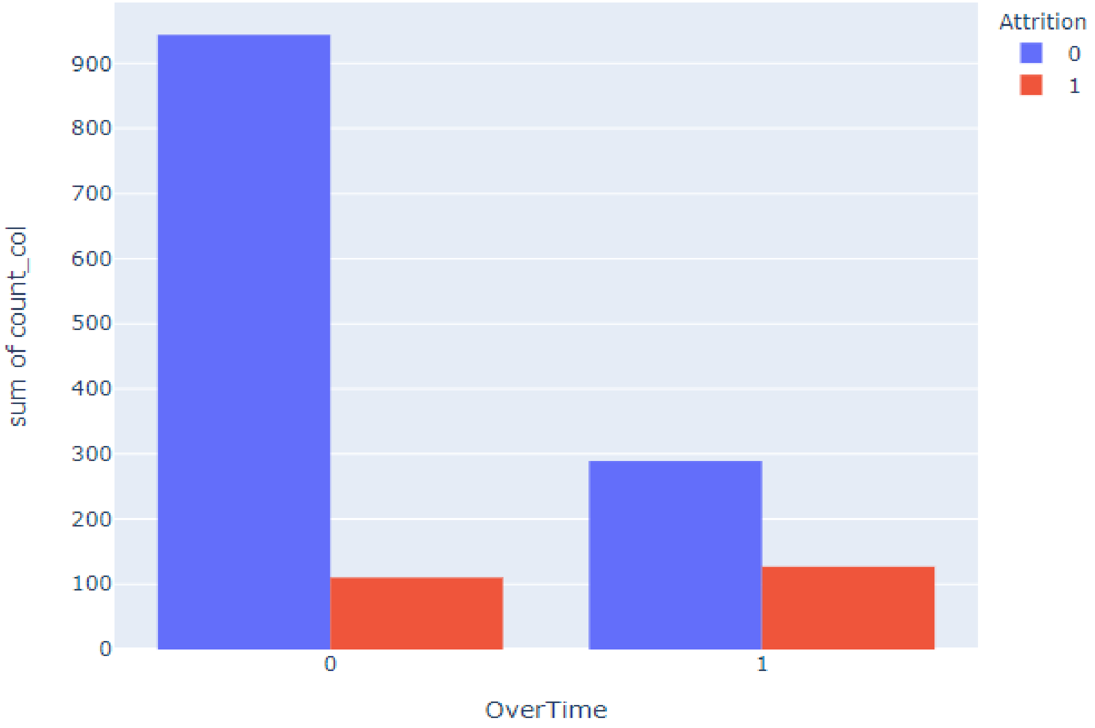
In the data set, we do not have a characteristic that qualifies the relationship with the boss, but we do have a column that quantifies the years with the current manager.



We can observe that employees who resign have less time with their manager than employees who keep their jobs

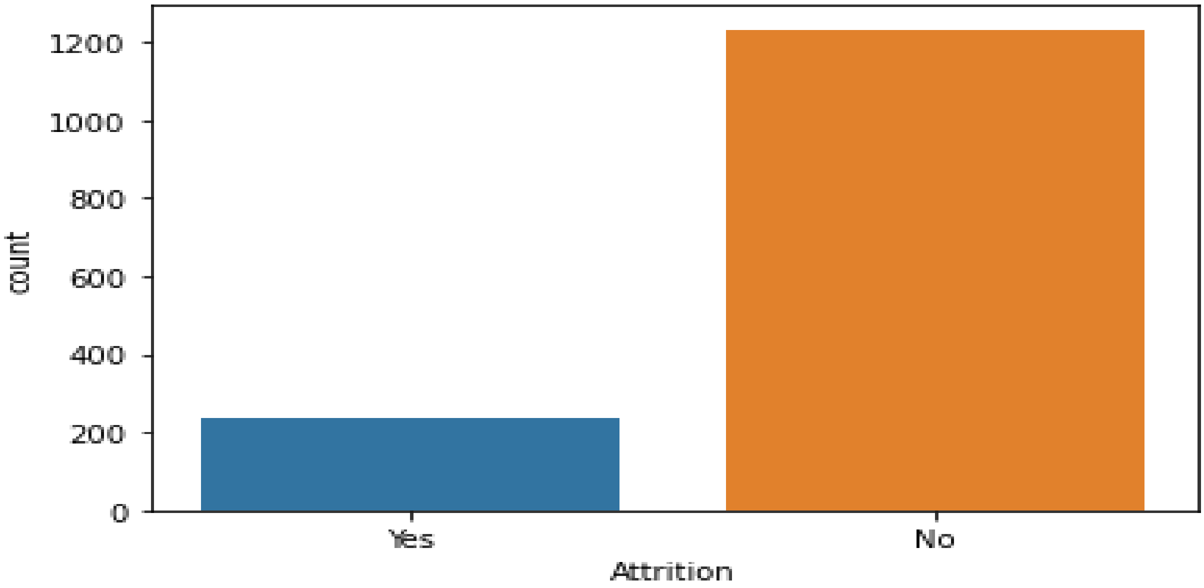
The emotional burnout that the job can generate is also a major factor in the employee quitting no matter what their salary and position in their job.

Burnout can be generated by overtime, so it is very important to find out what the relationship is between overtime and quitting.

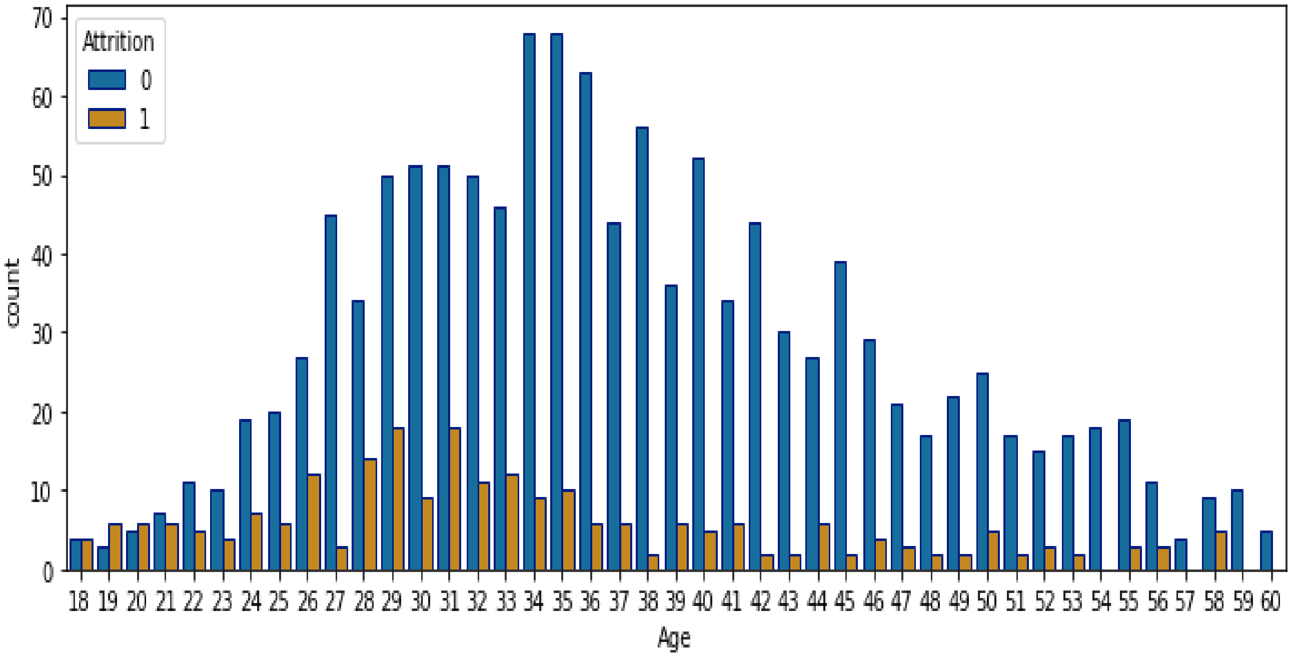


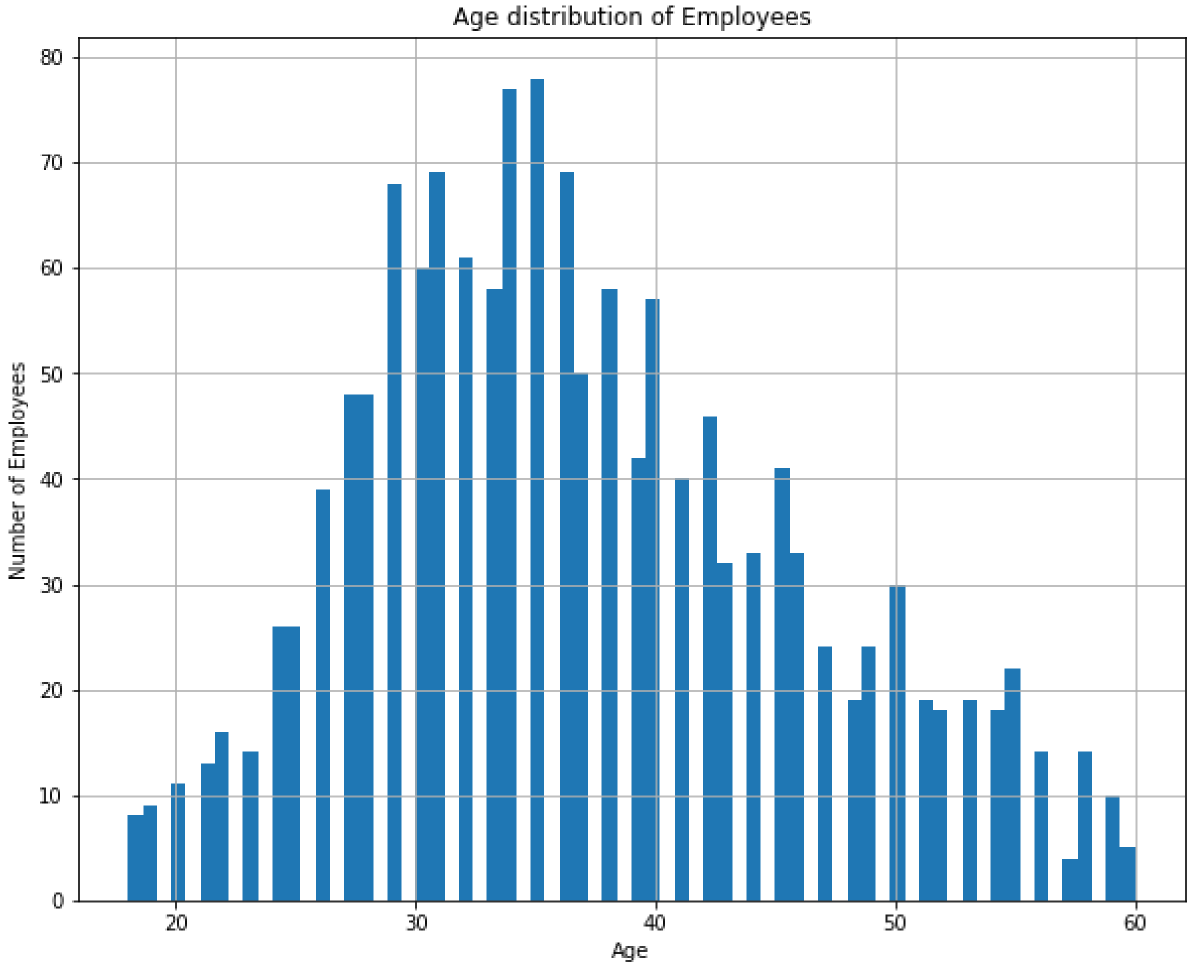
It is a fact that the number of employees who quit is higher when the employee works overtime

Exploratory Data Analysis



As per the data and we can observe that attrition rate means the people who quitting the jobs are around 20%





A short analysis reveals the following key points:

- For people who leave the company (on average):
    - They are between 33 years to 36 years
    - They live further from their work: 11km
    - Less satisfaction with the work environment: 2
    - Lower level of work: 1
    - Less satisfaction with work: 2
    - Lower monthly salary: $ 4800.00
    - Work more overtime: 0.5
    - Less years in the company: 5
    - Fewer years in current position: 2
    - Fewer years with current manager: 2

CHAPTER 2: Feature Engineering and Feature Selection

Data Cleaning

Missing Value

- After missing value treatment there where no missing in the data

To drop

- Employee count: values have the same value.
- Over18: all values have the same value.
- Standard hours: all values have the same value.
- Employee number: irrelevant variable, it is only an employee identifier.

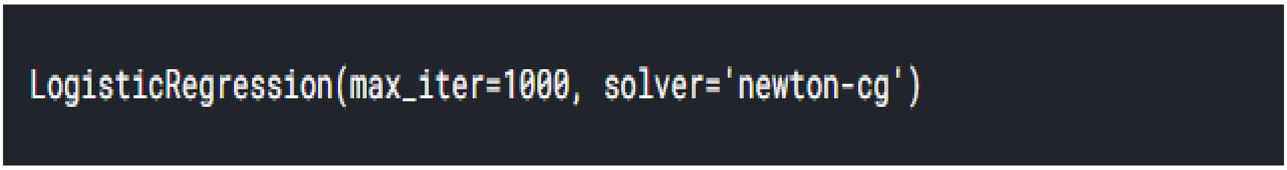
About daily rate, hourly rate, and monthly rate

- Monthly rate is the internal charge out rate which will be used to calculate the cost of each employee monthly, in general, the monthly rate will cover salary, social insurance, administration, logistics, overhead etc.
- Hourly rate and daily rate. These are not considered because the standard hours for every employee are 80 hours.

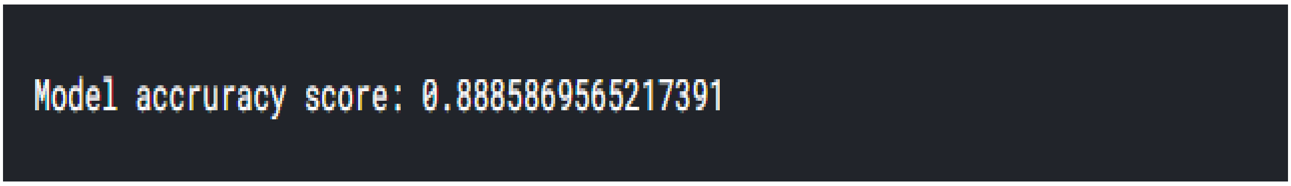
I decided to drop these three variables and keep only with "Monthly income" that is the total salary.

CHAPTER 3: FITTING MODELS TO DATA

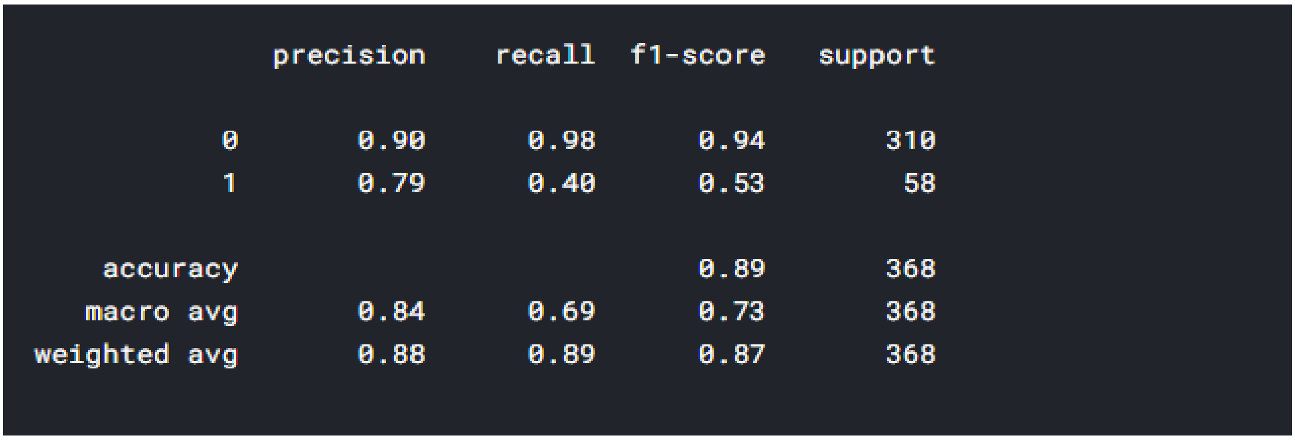
Linear regression



```
LogisticRegression(max_iter=1000, solver='newton-cg')
```



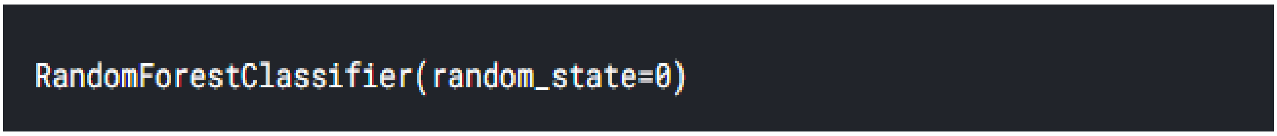
```
Model accruracy score: 0.8885869565217391
```



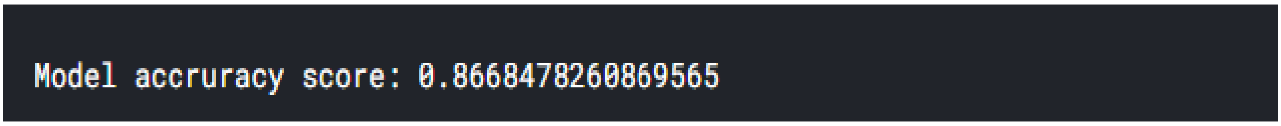
```
              precision    recall  f1-score   support

           0       0.90      0.98      0.94       310
           1       0.79      0.40      0.53        58

    accuracy                           0.89       368
   macro avg       0.84      0.69      0.73       368
weighted avg       0.88      0.89      0.87       368
```

We can see that the model predicts quite well the "none quite employees" (94% accuracy) but it doesn't predict as well the "quite employees" (53% accuracy).

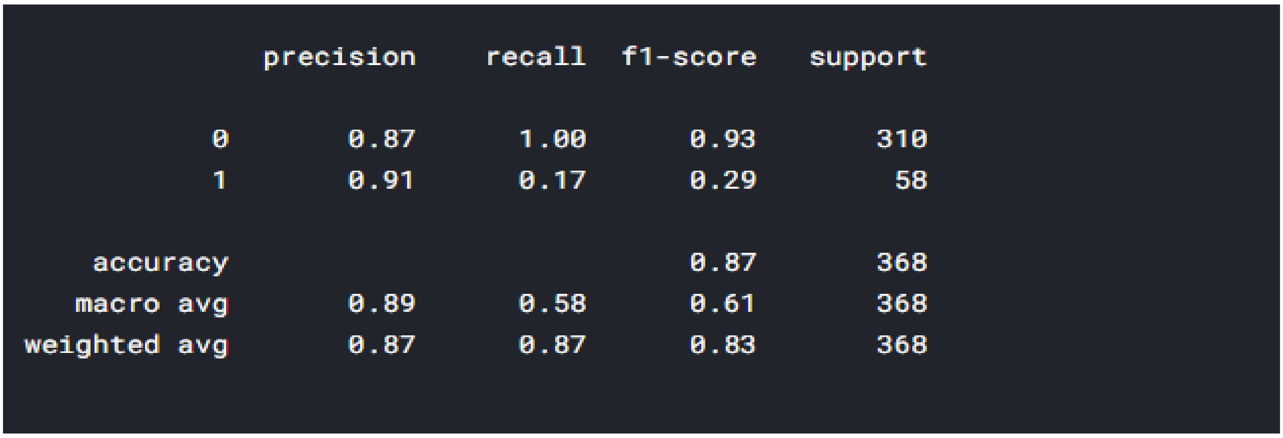
Random forest



```
RandomForestClassifier(random_state=0)
```



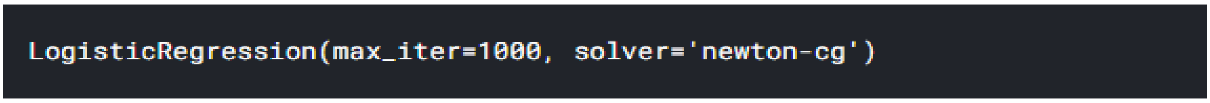
```
Model accruracy score: 0.8668478260869565
```



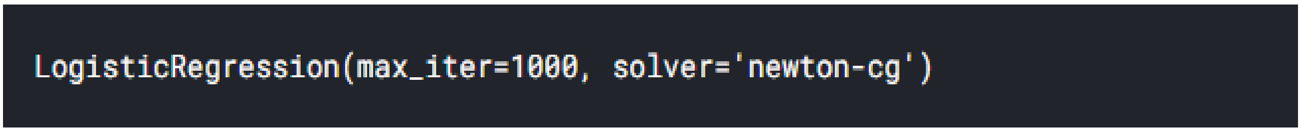
```
              precision    recall  f1-score   support

           0       0.87      1.00      0.93       310
           1       0.91      0.17      0.29        58

    accuracy                           0.87       368
   macro avg       0.89      0.58      0.61       368
weighted avg       0.87      0.87      0.83       368
```

The model predicts quite well the "none quitting employees" (93% accuracy) but it has a poor prediction of "quitting employees" (29% accuracy)

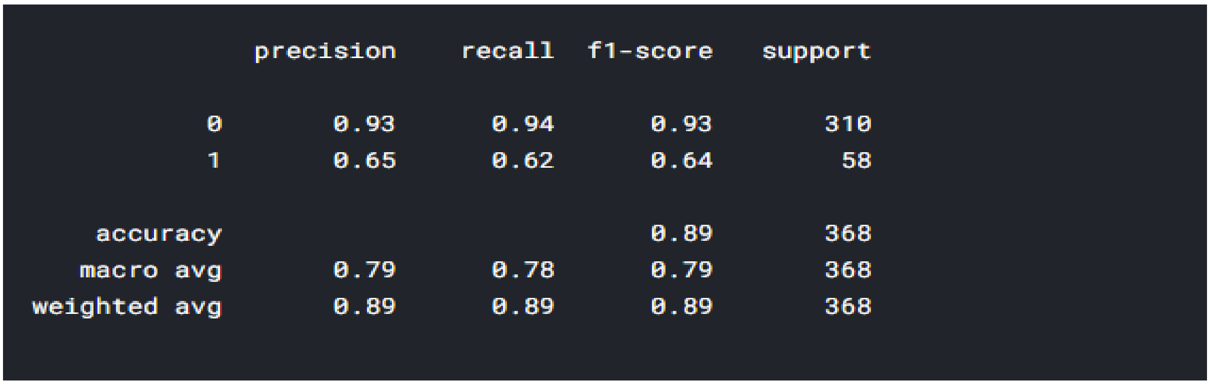
Logistic with SMOTE Data



```
LogisticRegression(max_iter=1000, solver='newton-cg')
```



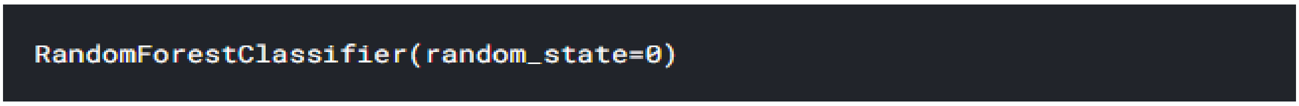
```
LogisticRegression(max_iter=1000, solver='newton-cg')
```



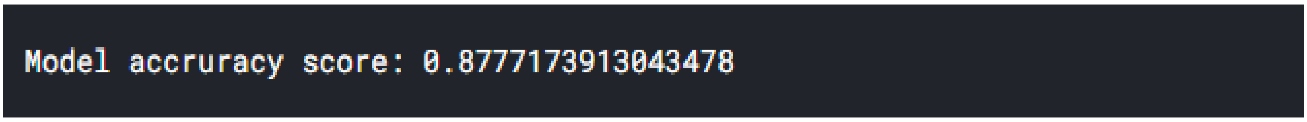
```
              precision    recall  f1-score   support

           0       0.93      0.94      0.93       310
           1       0.65      0.62      0.64        58

    accuracy                           0.89       368
   macro avg       0.79      0.78      0.79       368
weighted avg       0.89      0.89      0.89       368
```

With the SMOTE technique it is possible to get a better precision in the attrition cases (65 %)

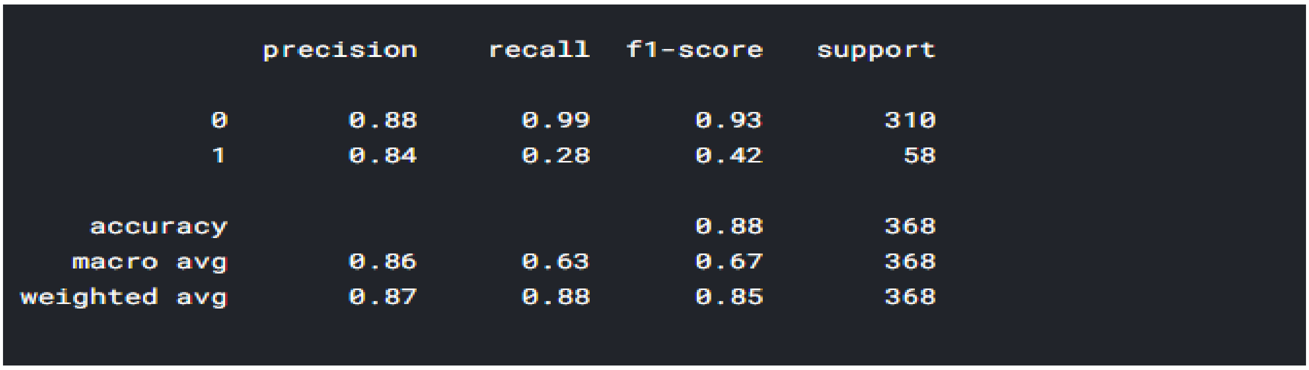
Random forest with SMOTE



```
RandomForestClassifier(random_state=0)
```



```
Model accruracy score: 0.8777173913043478
```



```
              precision    recall  f1-score   support

           0       0.88      0.99      0.93       310
           1       0.84      0.28      0.42        58

    accuracy                           0.88       368
   macro avg       0.86      0.63      0.67       368
weighted avg       0.87      0.88      0.85       368
```

In case of Random forest classifier, we have a better prediction of the "quite cases" but it doesn't do better than logistic regression model

CHAPTER 4: Conclusion and Recommendation

While non-competitive salary, poor work environment or bad relationship with the boss may be reasons for a worker to quit, these are not sufficient reasons for an employee to resign. Labour resignation is caused by a combination of multiple factors that may or may not be part of the characteristics of this dataset, however, it must be taken in consideration that each company will present diverse factors and ways of qualifying the worker, so this data stands in general to get the solution but still exactly may vary

About the model

Logistic regression proved to be a good tool to classify and predict which employees will not quit, however, the unbalance of the data set does not help to predict which employees will quit. To compensate for this, the SMOTE technique was used to generate synthetic data to compensate for the lack data from employees who quit. So recommending to use this tool carefully because it generates synthetic data around a cluster, which is not always good.