Corporate Employee Attrition Analysis

A PROJECT REPORT

Submitted by

Team ID : PNT2022TMID27015

Team Leader : NITHEESH RAAJ R M (310819104709)

Team member : THILAK C (310819104701)

Team member : RAGURAM R (310819104712)

Team member : HEMACHANDER S (310819104714)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

JEPPIAAR ENGINEERING COLLEGE

ANNA UNIVERSITY CHENNAI 600025

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The Nalaiya Theren opportunity I had with IBM in the Data Analytics domain was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual for being a part of it. I am also grateful to all the professionals who led me though this internship period.

Bearing in mind previous I am using this opportunity to express my deepest gratitude and special thanks to My Principal who helped me in spite of being extraordinarily busy with his duties, allowing me to carry out my internship at the esteemed organization.

I express my deepest thanks to my industry mentor Mr. Shanawaz Anwar for taking part in useful decision & giving necessary advices and guidance and arranged all facilities to make internship easier. I choose this moment to acknowledge his contribution gratefully.

It is my radiant sentiment to place on record my best regards, deepest sense of gratitude to our College faculty mentor, MADHURIKKHA S Assistant Professor, Department of CSE for his careful and precious guidance which were extremely valuable for my study both theoretically and practically.

I perceive as this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives.

Sincerely,

Team Leader : NITHEESH RAAJ R M

Team member : THILAK C

Team member : RAGURAM R

Team member : HEMACHANDER S

OBJECTIVE

- The objective of this project is to predict the attrition rate for each employee, to find who's more likely to leave the organisation.

- It will help organization to find ways to prevent attrition or plan in advance the hiring of the new candidate.

- Attrition proves to be a costly and time-consuming problem for the organization and it also leads to the loss of probability.

- The scope of the project extends to companies to all industries

DESCRIPTION OF PROJECT

- In this Project, it is required to clean and sanitize the dataset. Then, train the dataset to predict the attrition rate of the employees in an organization

METHODOLOGY

1. Business Understanding:

Before solving the problem in the Business domain it needs to be understood properly. Business understanding forms a concrete base, which further leads to easy resolution of queries. We should have the clarity of what is the exact problem we are going to solve.

2.Analytic Understanding:

Based on the above business understanding one should decide the analytical approach to follow. The approaches can be of 4 types: Descriptive approach (current status and information provided), Diagnostic approach (A.K.A statistical analysis, what is happening and why it is happening), Predictive approach (it forecasts on the trends or future events probability) and Prescriptive approach (how the problem should be solved actually).

3.Data Requirements:

The above chosen analytical method indicates the necessary data content, formats and sources to be gathered. During the process of data requirements, one should find the answers for questions like ‘what’, ‘where’, ‘when’, ‘why’, ‘how’ & ‘who’.

4.Data Collection:

Data collected can be obtained in any random format. So, according to the approach chosen and the output to be obtained, the data collected should be validated. Thus, if required one can gather more data or discard the irrelevant data.

5.Data Understanding:

Data understanding answers the question “Is the data collected representative of the problem to be solved?”. Descriptive statistics calculates the measures applied over data to access the content and quality of matter. This step may lead to reverting the back to the previous step for correction.

6.Data Preparation:

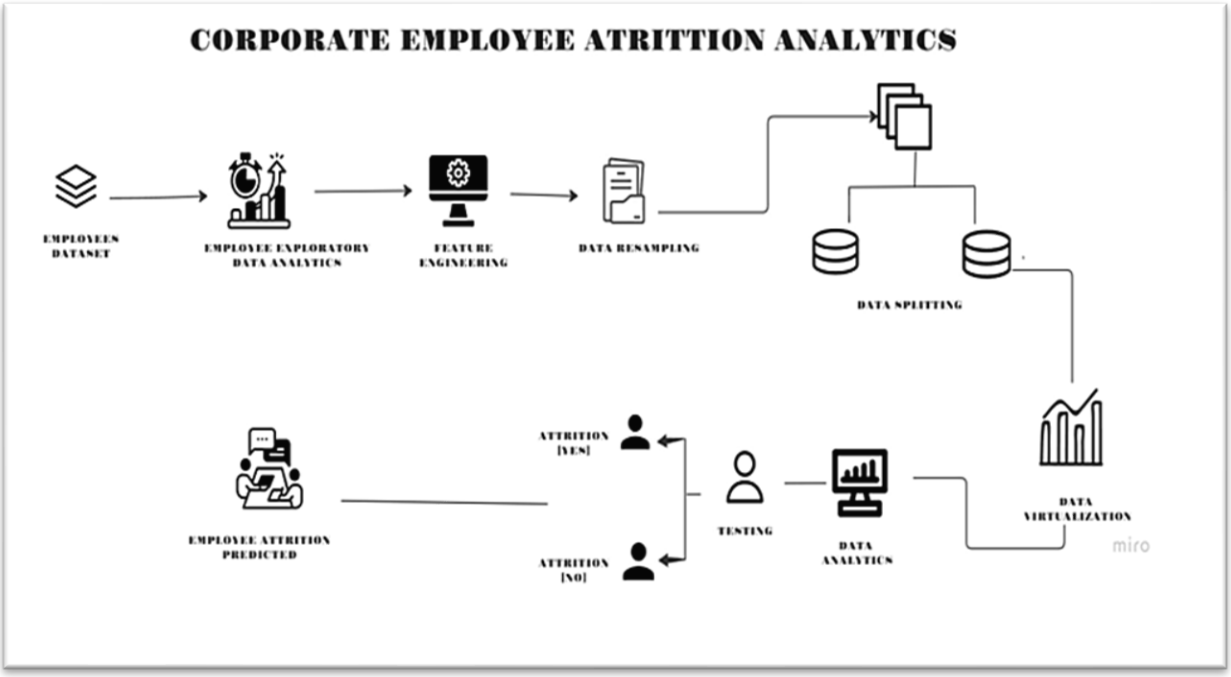
Let’s understand this by connecting this concept with two analogies. One is to wash freshly picked vegetables and second is only taking the wanted items to eat in the plate during the buffet. Washing of vegetables indicates the removal of dirt i.e. unwanted materials from the data. Here noise removal is done. Taking only eatable items in the plate is, if we don’t need specific data then we should not consider it for further process. This whole process includes transformation, normalization etc.

7.Modelling:

Modelling decides whether the data prepared for processing is appropriate or requires more finishing and seasoning. This phase focuses on the building of predictive/descriptive models.

8.Evaluation:

Model evaluation is done during model development. It checks for the quality of the model to be assessed and also if it meets the business requirements. It undergoes diagnostic measure phase (the model works as intended and where are modifications required) and statistical significance testing phase (ensures about proper data handling and interpretation).



9.Deployment:

As the model is effectively evaluated it is made ready for deployment in the business market. Deployment phase checks how much the model can withstand in the external environment and perform superiorly as compared to others.

10.Feedback:

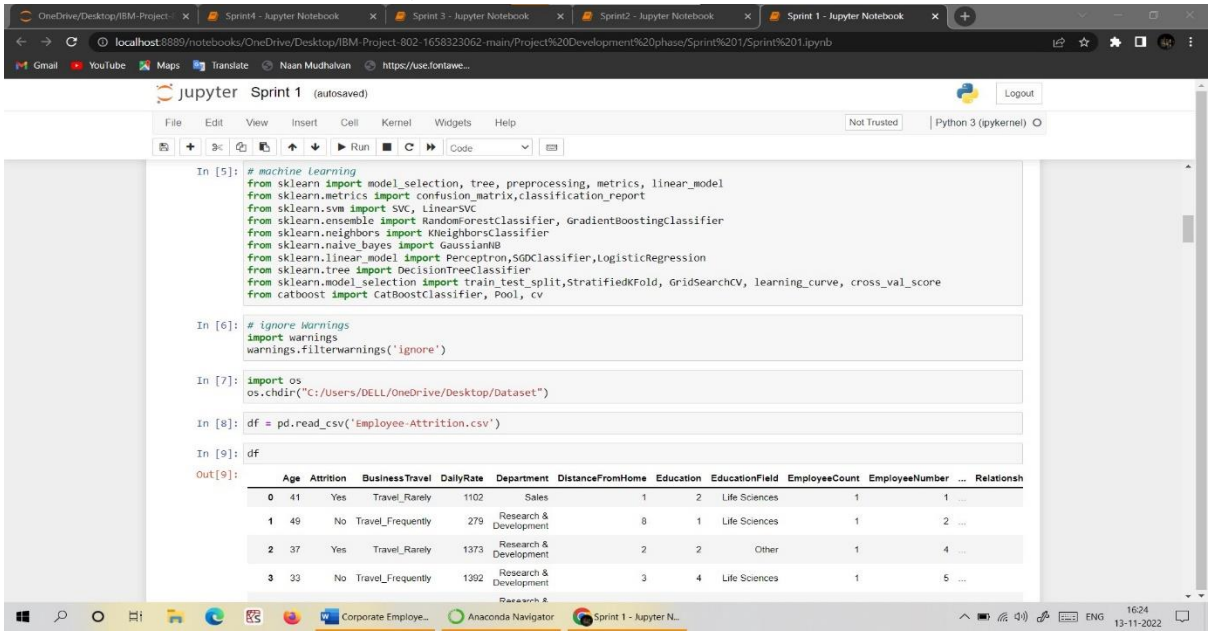
Feedback is the necessary purpose which helps in refining the model and accessing its performance and impact. Steps involved in feedback define the review process, track the record, measure effectiveness and review with refining.

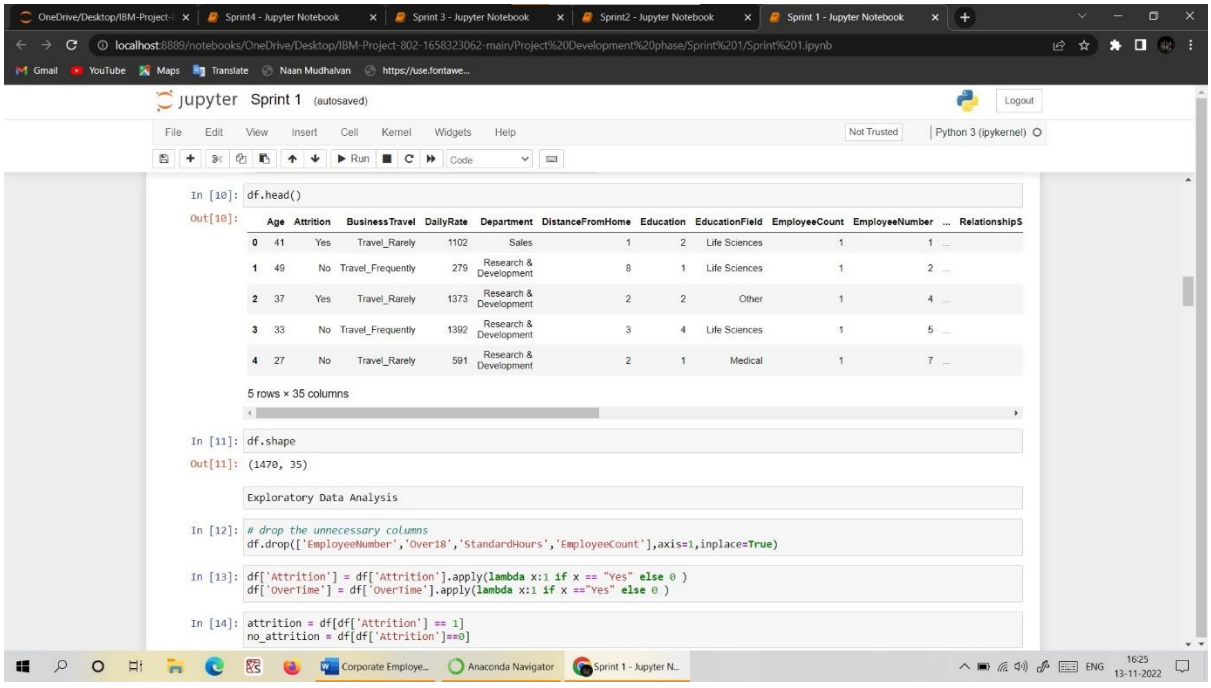
ASSUMPTIONS

1. Each possible sample has assigned to it a known probability of selection.
2. We select one of the samples by a random process in which each sample receives its appropriate probability of being selected.
3. The method for computing the estimate must lead to a unique estimate for any specific sample.
4. Homoscedasticity: The variance of residual is the same for any value of X.
5. Independence: Observations are independent of each other.
6. Normality: For any fixed value of X, Y is normally distributed.
7. Multicollinearity: There should be no or little multicollinearity.
8. No auto-correlation

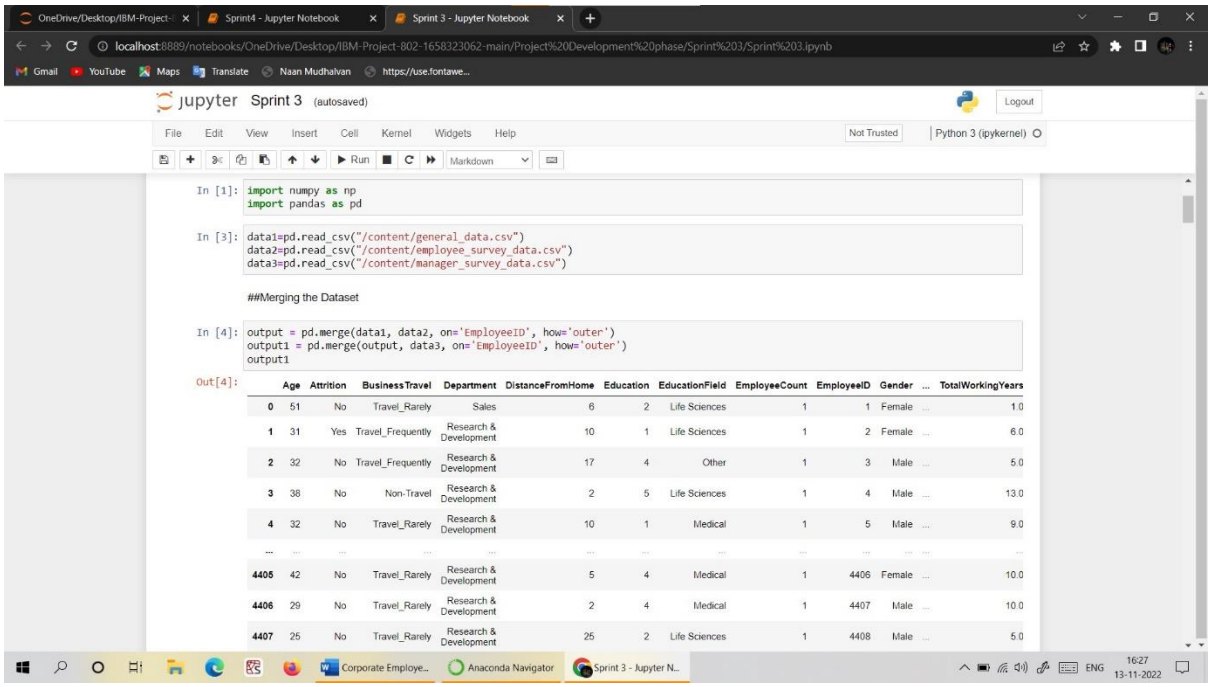
VISUALIZATIONS

This figure shows the data frame (which isn't cleaned and sanitized as it has lots of null values)

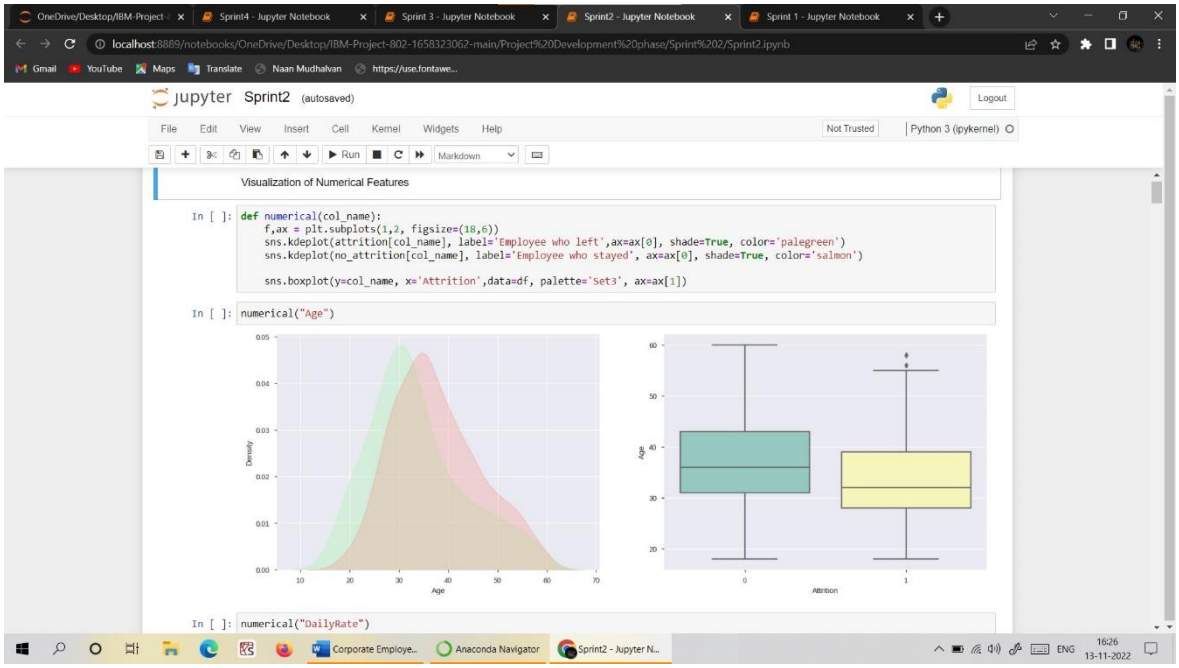


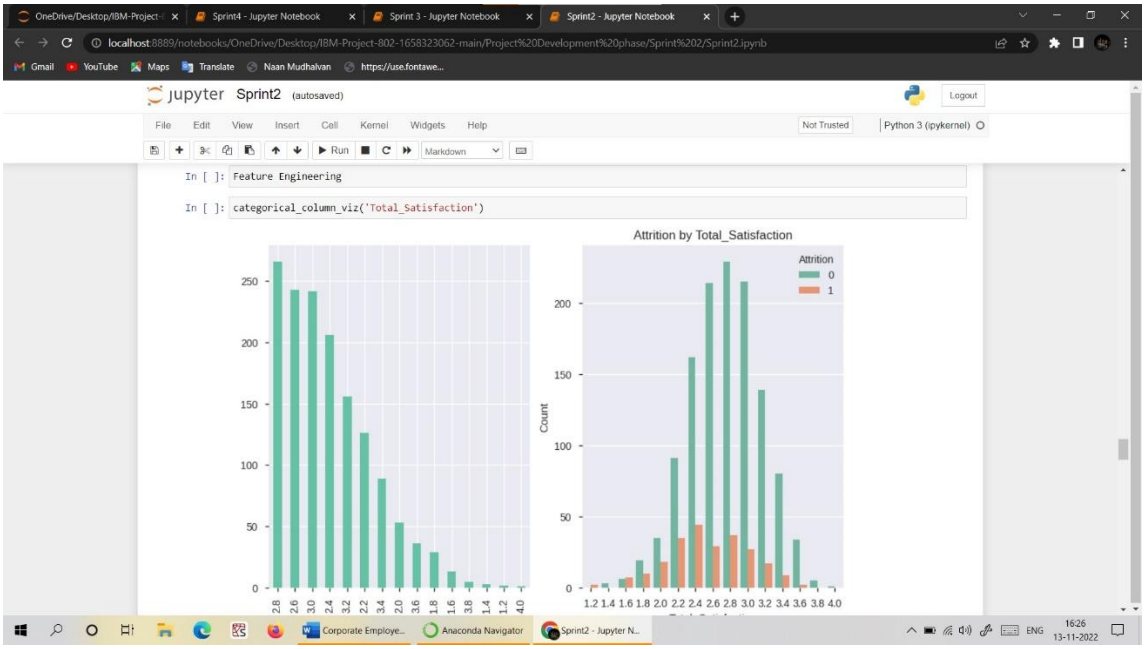


The below figure shows the dataset after removing unnecessary columns and the rows containing missing values and reordering the same.

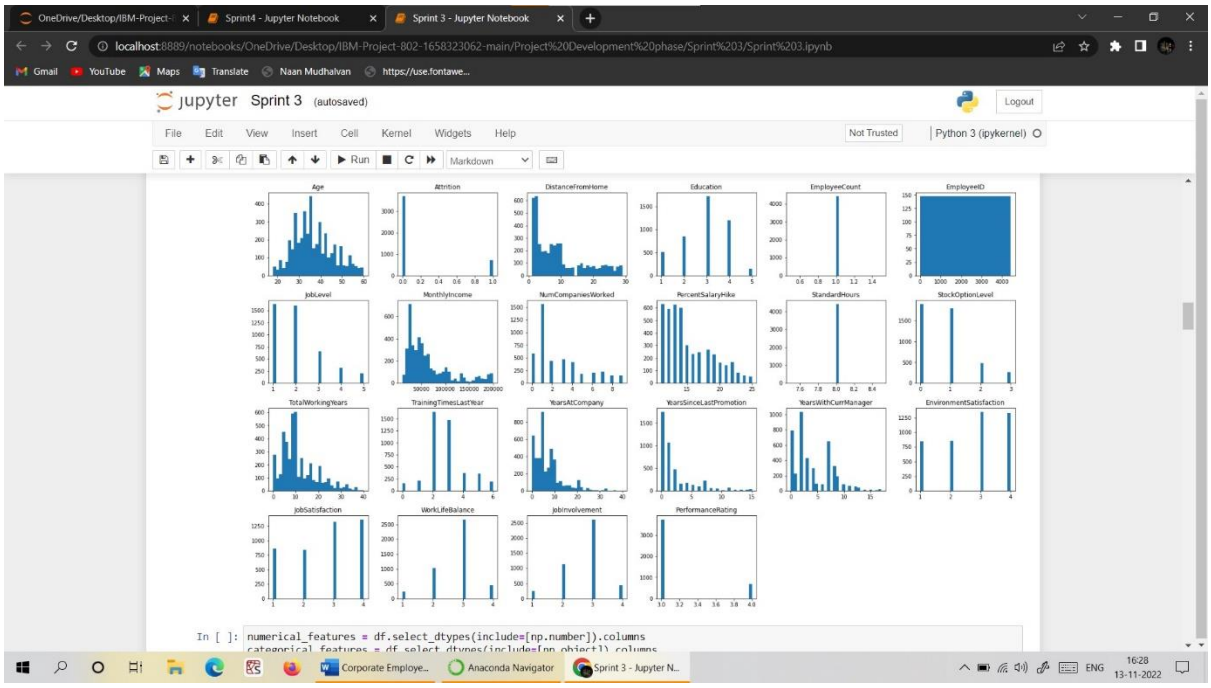


This figure shows the distribution of target variables

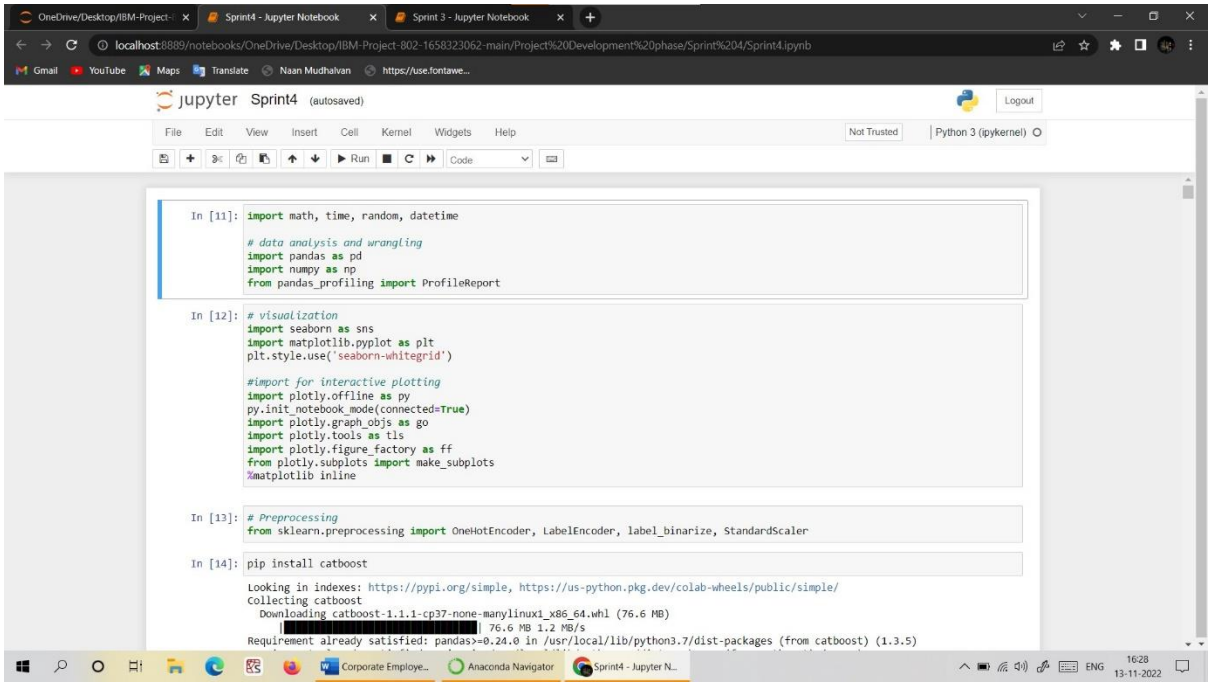




The visualisations of the numerical variables is



Similarly of the non numerical variables



And then, we train the model and test for unknown data to predict the attrition rate. And the screenshot is attached herewith.

```
# Logistic Regression
start_time = time.time()
train_pred_log, acc_log, acc_cv_log = fit_ml_algo(LogisticRegression(), X_train,y_train, 10)
log_time = (time.time() - start_time)
print("Accuracy: %s" % acc_log)
print("Accuracy CV 10-Fold: %s" % acc_cv_log)
print("Running Time: %s" % datetime.timedelta(seconds=log_time))
```

```
Accuracy: 89.8
Accuracy CV 10-Fold: 88.63
Running Time: 0:00:01.753627
```

```
# SVC
start_time = time.time()
train_pred_svc, acc_svc, acc_cv_svc = fit_ml_algo(SVC(),X_train,y_train,10)
svc_time = (time.time() - start_time)
print("Accuracy: %s" % acc_svc)
print("Accuracy CV 10-Fold: %s" % acc_cv_svc)
print("Running Time: %s" % datetime.timedelta(seconds=svc_time))
```

```
Accuracy: 88.53
Accuracy CV 10-Fold: 85.91
Running Time: 0:00:00.497278
```

```
# Linear SVC
start_time = time.time()
train_pred_svc, acc_linear_svc, acc_cv_linear_svc = fit_ml_algo(LinearSVC(),X_train, y_train,10)
linear_svc_time = (time.time() - start_time)
print("Accuracy: %s" % acc_linear_svc)
print("Accuracy CV 10-Fold: %s" % acc_cv_linear_svc)
print("Running Time: %s" % datetime.timedelta(seconds=linear_svc_time))
```

```
Accuracy: 89.89
Accuracy CV 10-Fold: 88.73
Running Time: 0:00:01.055932
```

REFLECTIONS ON THE PROJECT

Our Nalaiya Theren project in collaboration with IBM and Tamil Nadu government has been the most rewarding and learning experiences we have had. With such empathetic, compassionate and supportive mentors, this experience has helped me achieve my goal of completing my project . Because of the techniques we learned not only from my mentors and professors but from internet and books too.

We are confident that we will continue to grow and develop professionally and in my personal endeavours. Within my internship, there were two distinct learning experiences that stand out to me as the most influential aspects of my development this semester: community involvement in discussion forum and self-learning.

Throughout my project experience, we were able to develop and foster a truly positive and compassionate learning cum implementation environment, all through the support and mentorship of our mentors.

Through the application of time management, organization, discipline and consistent practice, our self exploration and learning skills improved greatly. Additionally, my development both with the project we were given with and planning and implementing the same directly impacted our academic gain.

We are confident in our growth and development. We would not have the knowledge or skills we have today if it were not for our project experience with the industry mentor, college mentors and fellow interns.

CONCLUSION

On the whole, this project was a useful experience. We have gained new knowledge and skills we achieved several of my learning goals. We got insight into professional practice. We learned the different facets of working .

We experienced that self exploration, as in many organisations, is an important factor for the progress of projects. Related to our study we learned more about employee attrition rate prediction and the various approaches and algorithms to achieve the same.

There is still a lot to discover and to improve. The methods used at the moment are still not standardized and a consistent method is in development.

Furthermore we have experienced that it is of importance of each strategy and how other one is better than the current algorithm and in which application. we found that the internship is not one sided, but it is a way of sharing knowledge, ideas and opinions and implementing the same to get results.

The internship was also good to find out what our strengths and weaknesses are. This helped me to define what skills and knowledge.

We believe that our time spent in learning and surfing regarding various algorithms and the mathematics behind was well worth it and contributed to finding an acceptable solution to build a model and predict the employee’s attrition rate. Two main things that we’ve learned the importance of time-management skills and self-motivation.

At last this project has gave us new insights and motivation to pursue a career in machine learning domain.

Link to code and executable file

https://github.com/IBM-EPBL/IBM-Project-802-1658323062.git