****IBM HR Employee Attrition Rate**

**ABSTRACT:** Employees are the most valuable assets to any organization. They are the fulcrum and laying foundation of an organisation. It is they who add value to the organization in terms of quantity as well as quality. To find, attract, develop and retain the right talent is a critical task for the management. Therefore, it is indispensable to maintain a permanent and promising workforce which over the years has become a tough task for employers and thereby resulting in increased attrition. This comprehensive analysis on the IBM dataset aims to make a detailed study on attrition, comprehend to various interesting features ,to forecast attrition so employers are prepared well in advance and most importantly implement ways to retract rates of attrition in the future.

Presented by ,

Group 8 Members :

Sagar Shee, Barasa Sarma, Modak Shivananda , Shaik Tahir, Shawin Pradhan

**Acknowledgement**

We would like to take this opportunity to express our profound gratitude and deep regards to our mentor guide Mr. Shashank Prakash Shirude for his exemplary guidance, mentoring and constant encouragement throughout the course of this Project.

We would like to thank all our peers, through whom the learning grew exponentially.

**GREAT LAKES INSTITUTE OF MANAGEMENT**

**CERTIFICATE**

This is to certify that this project report entitled **“IBM HR EMPLOYEE ATTRITION RATE”** submitted to **Great Learning, Bangalore**, is a bonafide record of work done by **Barasa Sarma, Sagar Shee, Shawin Pradhan,**

**Modak Shivananda and Tahir Shaik** in partial fulfilment of the requirements for the award of Post Graduate Program in Data Science and Engineering at Great Learning, Bangalore. It is a work carried out by them under my supervision and guidance.

To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other University/Institute for the award of any Degree or Diploma.

Date**: Mr. Shashank Prakash Shirude**

**Table of Contents:**

**1 Introduction………………………………………………………………… 3**

1.1. What Is Attrition?..........................................................................................3

1.2 Industry Review……………………………………………………………3

1.3 Background Research...................................................................................4

1.4. Need of the Study ........................................................................................4

1.5. Scope The Study...........................................................................................4

1.6 Objective of The Study…………………………………………………….5

**2. Dataset and Domain ………..........................................................................6**

2.1 Data Dictionary ……………….....................................................................7

2.2 Data Attributes………...................................................................................7

2.3 Irrelevant Columns ........................................................................................8

2.4 Project Justification…………………………………………………………9

**3. Statistical Analysis………………………………………………………….10**

3.1 Chi-square Test..……………………………………………………………10

3.2 ANOVA…………………………………………………………………... 11

3.3 Type 1 and Type 2 error …………………………………………………...12

**4. Exploratory Data Analysis…………………………………………………13**

4.1 Data Exploration……………………………………………………………13

4.2 Principal Component Analysis (PCA)……………………………………...31

**5 Feature Engineering………………………………………………………...33**

5.1 Outlier Detection …………………………………………………………..33

5.2 Data Transformation………………………………………………………..35

5.3 Scaling the Data…………………………………………………………….36

5.4 Feature Extraction………………………………………………………….37

5.5 Feature Selection …………………………………………………………..39

**6. Model Building .............................................................................................41**

6.1 Model Implementation (GAME PLAN) …………………………………...41

**6.2 Linear Models………………………………………………….……….…43**

6.2.1 Logistic Regression………………………………………………………46

6.2.2 Naïve Bayes …………………………………………………………...…49

6.3.1 Decision Tree…………………………………………………………….50

6.2.2 Random Forest…………………………………………………………...51

6.2.3. K-Nearest Neighbor……………………………………………………..52

6.2.4 Support Vector Machines………………………………………………...53

6.3 Ensemble Techniques………………………………………………………54

6.4 Comparison of all models….……………………………………………….56

**7. Limitations………………………………………………………………….60**

**8. Conclusion…...………………………………………………….……….…61**

**1. Introduction**

**1.1 What is “ATTRITION”?**

Attrition is defined as a gradual reduction of the size of workforce through normal means, such as retirement, resignation or death. This is normal in any business or industry. Attrition rate is defined as the rate of shrinkage in size or number.

* 1. **Industry Review:**
* Currently IBM artificial intelligence can predict which employees will leave a job with 95 percent accuracy.
* AI, which has replaced 30 percent of IBM’s HR staff, can help employees identify new skills training, education, job promotions and raises.
* IBM’s bet is that the future of work is one in which a machine understands the individual better than the HR individual can alone
* According to the Consumer Technology Association’s Future of Work survey, yet the tech industry is concerned that school systems and universities have not moved fast enough to adjust their curriculum to delve more into data science and machine learning. As a result, companies will struggle to fill jobs in software development, data analytics and engineering.
* IBM is investing $1 billion in initiatives like apprenticeships to train workers for what it calls “new collar” jobs.
* The “new collar” jobs could range from working at a call centre to developing apps or becoming a cyber-analyst at IBM after going through a P-TECH (Pathways in Technology Early College High School) program, which takes six years starting with high school and an associate’s degree

**1**.**3 Background Research:**

* Earlier top technology services companies across the world have the problem of high employee turnover in an industry that typically witnesses high attrition rates.
* However, with the emergence of predictive algorithms and tools that reads data in seconds and provide crucial insights and indicators into employee behaviour And the likes of IBM are investing heavily on such predictive analytical tools, which otherwise is wasted on hires that may not have been the best fit for the company in the first place.
* Targeted at the people who are the most productive and the most likely to stay, because you can save a lot of money on people who have a low probability of staying with you. IBM found that they could save millions of dollars by targeting our benefits and compensation for the people who are most productive and most likely to stay with IBM.
* Companies like IBM have, therefore, been forced to find new ways of retaining employees and have had to take a fresh look at traditional technology industry metrics. And the results are starting to show, as companies are now deploying a more strategic approach towards hiring and only investing on talent that is likely to help drive long-term growth.

Challenges at this point for IBM is to try to find some general models that will bring down the cost of doing this work, so that we can apply it to help people. In the past few years, companies have gathered socioeconomic data from incoming engineers such as the educational qualifications of parents and household incomes. Armed with such information, human resource (HR) departments are able to use algorithms and analytics in recruitment.

**1.4 Need Of The Study:**

The study was mainly undertaken to identify the level of employee’s attitude, the dissatisfaction factors they face in the organization and for what reasons they prefer to change their job. Once the levels of employee’s attitude are identified, it would be possible for the management to take necessary actions to reduce attrition level. Since they are considered as the backbone of any company, their progression and development will lead to exponential and consistent growth of companies in the longer run.

**1.5 Scope Of The Study:**

* To determine effect of attrition on the business.
* Determination of solutions to avoid or to control attrition.
* To understand the extent of job satisfaction among the employees.
* To suggest proper measures.
* This study helps the company to understand more on the attrition rate in the company,
* The study also helps to find the drawbacks of the current retention strategies.

**1.6 Objectives of The Study:**

* To foster a pattern during the stint of an employee as to when and what factors lead to attrition based on available facts and historical data.
* To analyse the parameters of a dissatisfied employee and adopt curated set of innovative ideas focusing on employee needs and desires.
* To consistently retain a low percentage of attrition to not only acquire top notch employees but discretely invite the interests of potential investors.
* To deploy an ascertained model which forecasts attrition beforehand and directly revamps a smooth decision-making process.
* To constitute state of the art retention policies and framework that would mutually benefit both the organization and employees while fixating on the reasons of attrition mentioned above.

**2. Dataset and Domain**

**2.1 Data Dictionary**

Data Rows: 1470 entries Data Columns :35 entries

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No** | **Feature Name** | **Data Type** | **Missing Value Count** |
| 1 | Age | Numerical | 0 |
| 2 | Attrition | Categorical | 0 |
| 3 | Business Travel | Categorical | 0 |
| 4 | Daily Rate | Numerical | 0 |
| 5 | Department | Categorical | 0 |
| 6 | Distance From Home | Numerical | 0 |
| 7 | Education | Numerical | 0 |
| 8 | Education Field | Categorical | 0 |
| 9 | Employee Count | Numerical | 0 |
| 10 | Employee Number | Numerical | 0 |
| 11 | Environment Satisfaction | Numerical | 0 |
| 12 | Gender | Categorical | 0 |
| 13 | Hourly Rate | Numerical | 0 |
| 14 | Job Involvement | Numerical | 0 |
| 15 | Job Level | Numerical | 0 |
| 16 | Job Role | Categorical | 0 |
| 17 | Job Satisfaction | Numerical | 0 |
| 18 | Marital Status | Categorical | 0 |
| 19 | Monthly Income | Numerical | 0 |
| 20 | Monthly Rate | Numerical | 0 |
| 21 | Num Companies Worked | Numerical | 0 |
| 22 | Over18 | Categorical | 0 |
| 23 | Overtime | Categorical | 0 |
| 24 | Percentage Salary Hike | Numerical | 0 |
| 25 | Performance Rating | Numerical | 0 |
| 26 | Relationship Satisfaction | Numerical | 0 |
| 27 | Standard Hours | Numerical | 0 |
| 28 | Stock Option Level | Numerical | 0 |
| 29 | Total Working Years | Numerical | 0 |
| 30 | Training Times Last Year | Numerical | 0 |
| 31 | Work Life Balance | Numerical | 0 |
| 32 | Years At Company | Numerical | 0 |
| 33 | Years In Current Role | Numerical | 0 |
| 34 | Years Since Last Promotion | Numerical | 0 |
| 35 | Years With Curr Manager | Numerical | 0 |

**The dataset contains :**

* 26 **Numerical Variables (Discrete:**26  **Continuous :**0 **)**
* 9 **Categorical Variables (Ordinal: Nominal:**9 **)**
* **The dataset has no missing values in any of the features.**

**2.2 Feature Attributes:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No** | **Features** | **Unique elements in each feature** | **Data Type** |
| 1 | Attrition | * Yes * No | Categorical |
| 2 | Business Travel | * Non-Travel * Travel-Frequently * Travel-Rarely | Categorical |
| 3 | Department | * Sales * Research & Development * HR | Categorical |
| 4 | Education | * 1(Middle School) * 2(High School) * 3(UG) * 4(PG) * 5(Ph.D.) | Numerical |
| 5 | Education Field | * 1(Human Resources) * 2(Life Sciences) * 3(Medical) * 4(Marketing) * 5(Technical Degree) * 6(Other) | Categorical |
| 6 | Environment Satisfaction | * 1(Low) * 2(Medium) * 3(High) * 4(Very High) | Numerical |
| 7 | Gender | * Male * Female | Categorical |
| 8 | Job Involvement | * 1(Low) * 2(Medium) * 3(High) * 4(Very High) | Numerical |
| 9 | Job Level | * 1(Junior) * 2(Intermediate) * 3(Senior) * 4(Vice-President) * 5(President) | Numerical |
| 11 | Job Satisfaction | * 1(Low) * 2(Medium) * 3(High) * 4(Very High) | Numerical |
| 12 | Marital Status | * Single * Married * Divorced | Categorical |
| 13 | Over18 | * Yes | Categorical |
| 14 | Overtime | * Yes * No | Categorical |
| 15 | Performance Rating | * 3(Excellent) * 4(Outstanding) | Numerical |
| 16 | Relationship Satisfaction | * 1(Low) * 2(Medium) * 3(High) * 4(Very High) | Numerical |
| 17 | Stock Option Level | * 0 * 1 * 2 * 3 | Numerical |
| 18 | Work Life Balance | * 1(Bad) * 2(Good) * 3(Better) * 4(Best) | Numerical |
| 10 | Job Role | * 1(Sales Executive) * 2( Research Scientist) * 3( Laboratory Technician) * 4( Manufacturing Director) * 5( Healthcare Representative) * 6( Manager) * 7( Sales Representative) * 8(Research Director) * 9(Human Resources) | Categorical |

**2.3 Irrelevant Columns:**

There are few features which are redundant and do not contribute in the prediction of the target variable. Hence we dropped them.

|  |  |
| --- | --- |
| **Features** | **Justification** |
| Standard Hours | 80 Standard Hours is same for all the records |
| Over18 | ‘YES’ is redundant to all records |
| Employee Count | Pointless feature with ‘1’ for all records |
| Employee Number | Using this feature as the index colum |

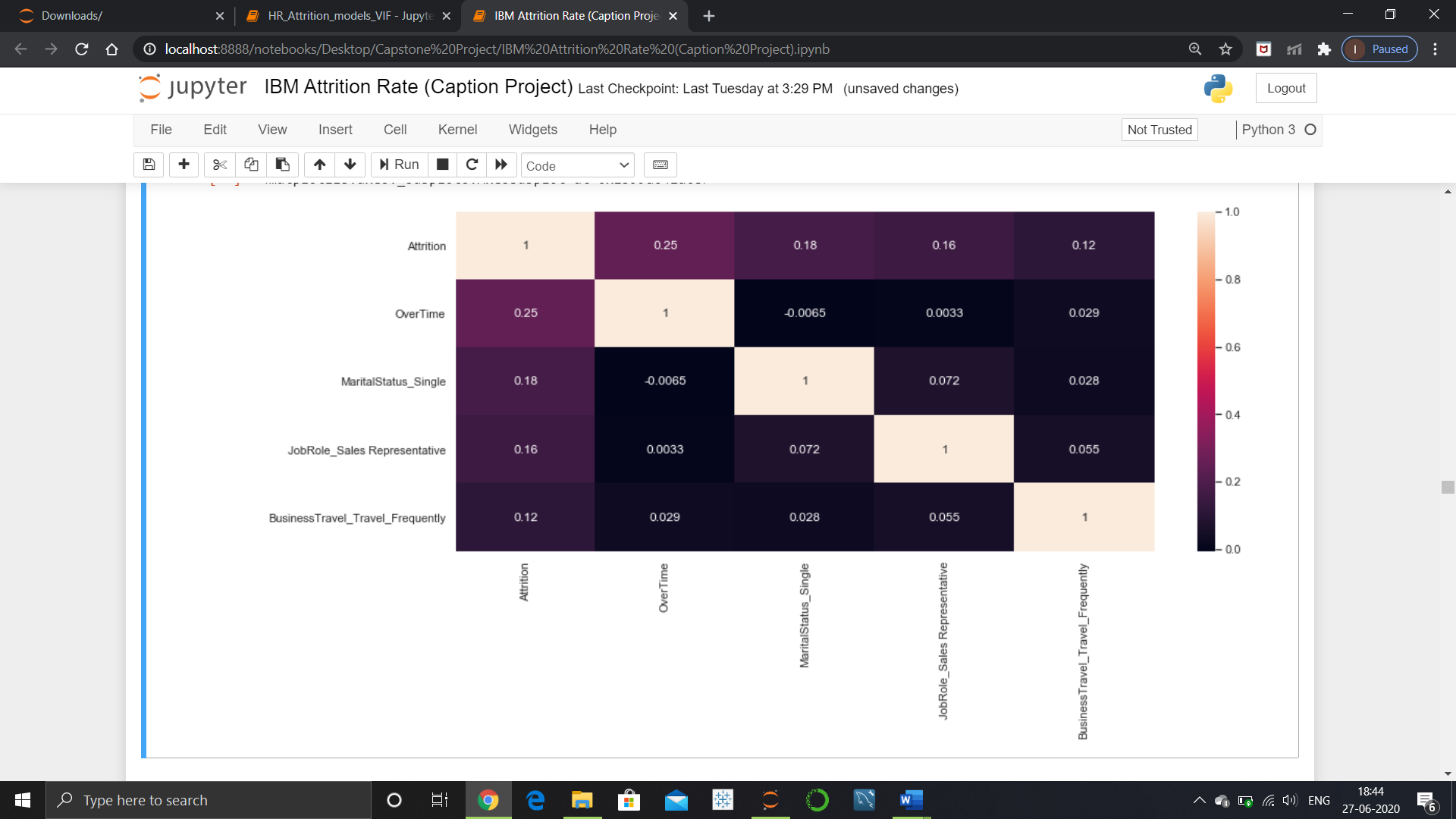
**2.4 Project Justification:**

1. This is a fictional dataset created by IBM data scientists and contains employee details which focuses primarily on HR analytics.
2. This is a classic classification problem and the dependant variable is **Attrition.**
3. We used all classification algorithms like Logistic Regression, Naïve Bayes, KNN, Decision Tree, Random Forest and Support Vector Machines and scaled intricate bagging and boosting techniques to increase the accuracy and performance of the model.

* **Project Complexity**

1. Complexity is substantial as the dataset is well crafted with no missing values and enough features to comprehensively analyse the pain points for attrition.
2. Feature extraction and feature engineering was minimal as most features were already pre-defined and on point.
3. Most numerical features were normally distributed therefore minimal data transformation techniques were required.

* **Project Outcome**

1. With a good prediction score companies will have a better understanding of their respective employees on a commercial level.
2. Companies can now formulate smarter retention systems and programs to maintain employee satisfaction.
3. Academically speaking, the dataset is fantastic to refresh graphical and visual representation of important features. Features like employee, environment and manger satisfaction levels are considered. Work life balance and job role are important factors to all employees and using these variables the company can weigh the probabilities of retention precisely.
4. Towards the end of the analyses we found that factors like overtime, marital status and distance from home are critical pain points and highly influence attrition rates.

**3. Statistical Analysis**

Questions??

\* Is gender influencing attrition?

\* Which job role has a higher signifance in predecting attrition?

\* Is Distance from home an important attribute?

To answer these intricate questions and to confidentally arrive at a solid conclusion , statistical analysis is performed on each feature.

\* H0=Feature = Attrition

\* H1=Feature!= Attrition (Explains the feature is having influence on attiriton)

to formulate a hypothesis with Gender ,

\* h0=mu male = mu female

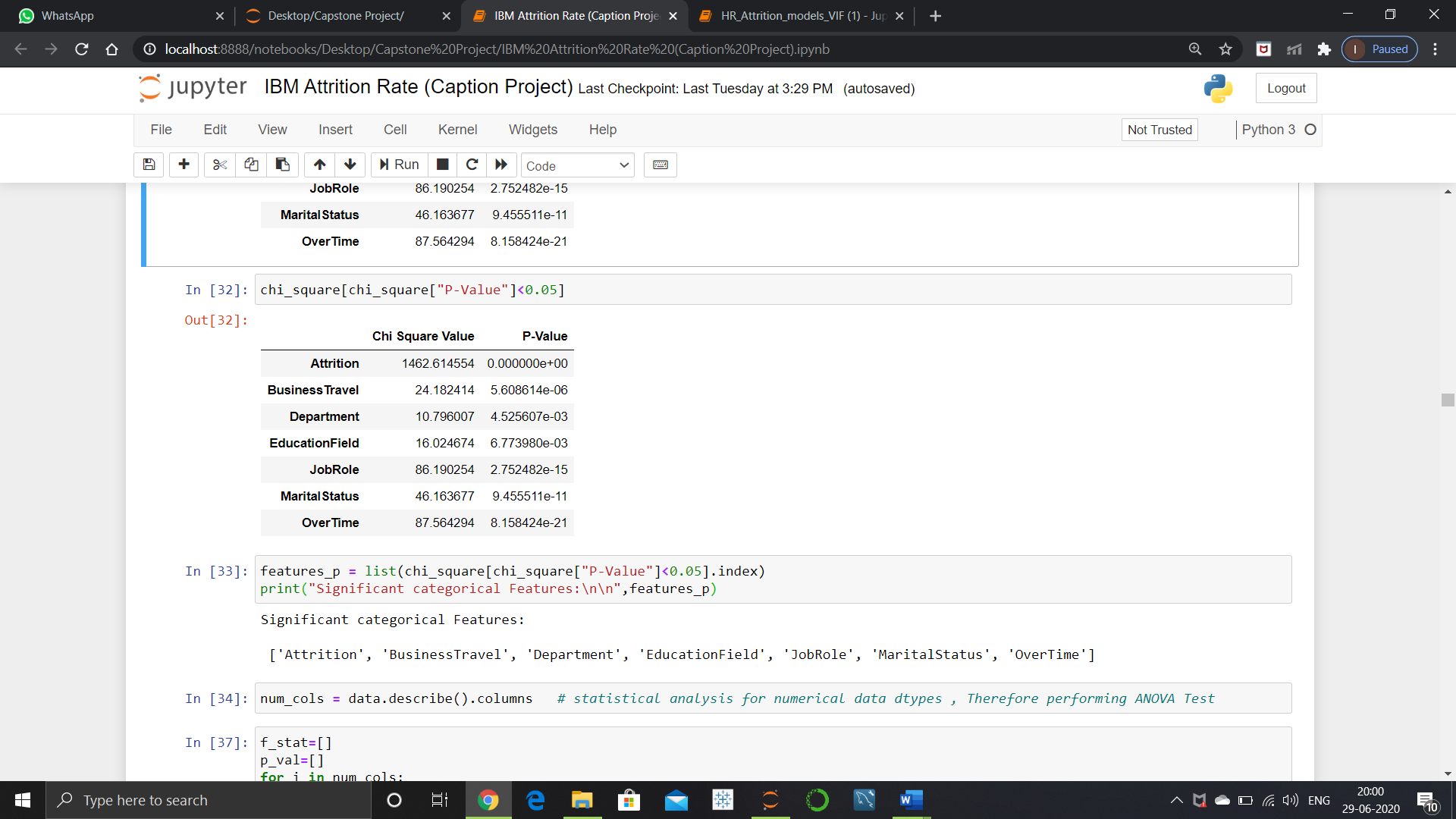
\* h1=mu male!= mu female

to formulate a hypothesis with Department ,

\* h0=mu Sales = mu Research & Development= mu Human Resources

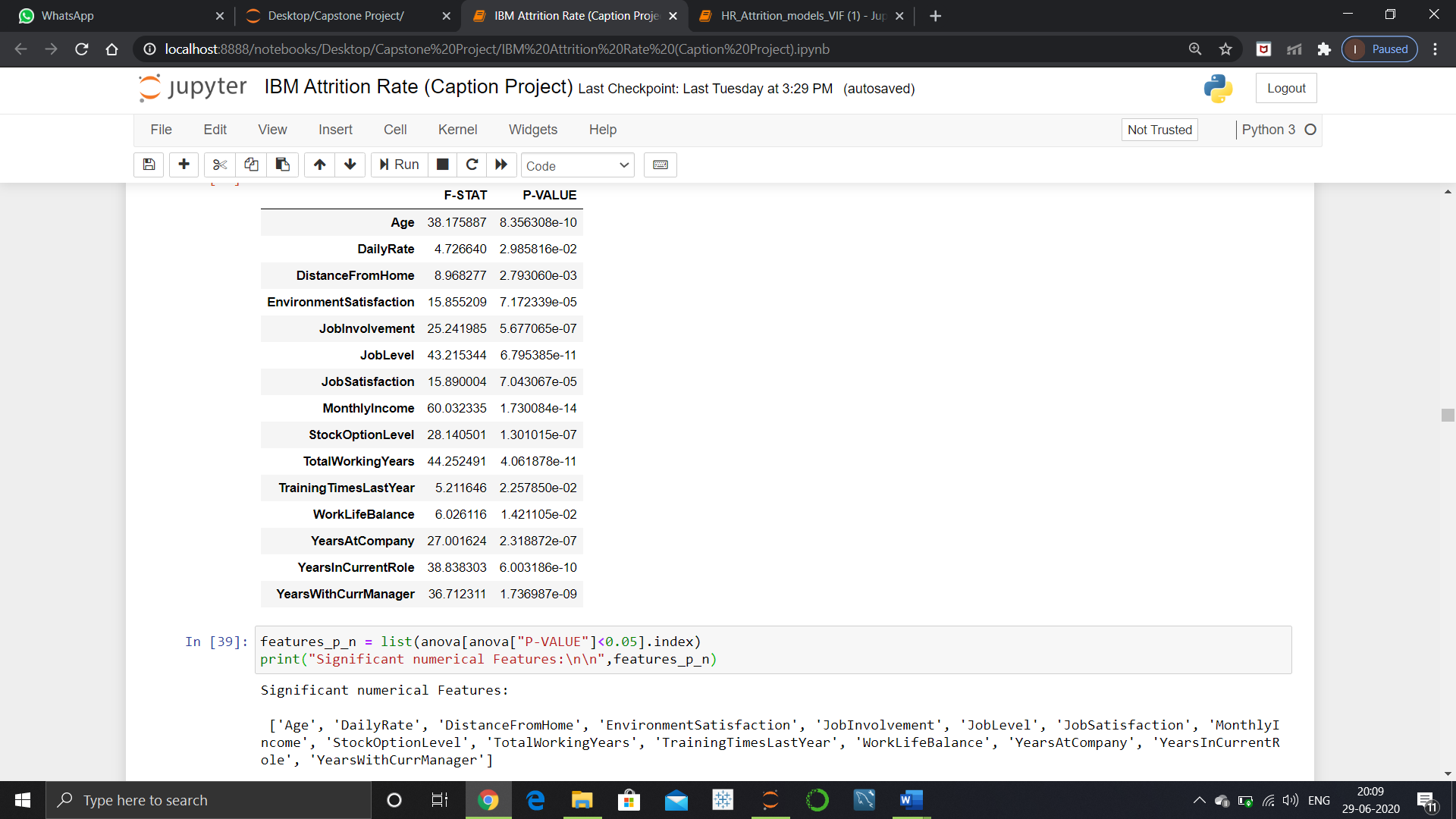
\* h1=mu Sales != mu Research & Development!= mu Human Resources

**3.1** **Chi-square Test**

Here, Chi-square Test is performed to evaluate the independence and importance of each feature with respect to attrition.These significant features have a P-value below 0.05 , due to which we reject the null hypothesis.

**3.2** **ANOVA**

Here, **ANOVA** is performed to understand what features highly influence rate of attrition, features here are of numerical data type.



* 1. **Type 1 and Type 2 Error**

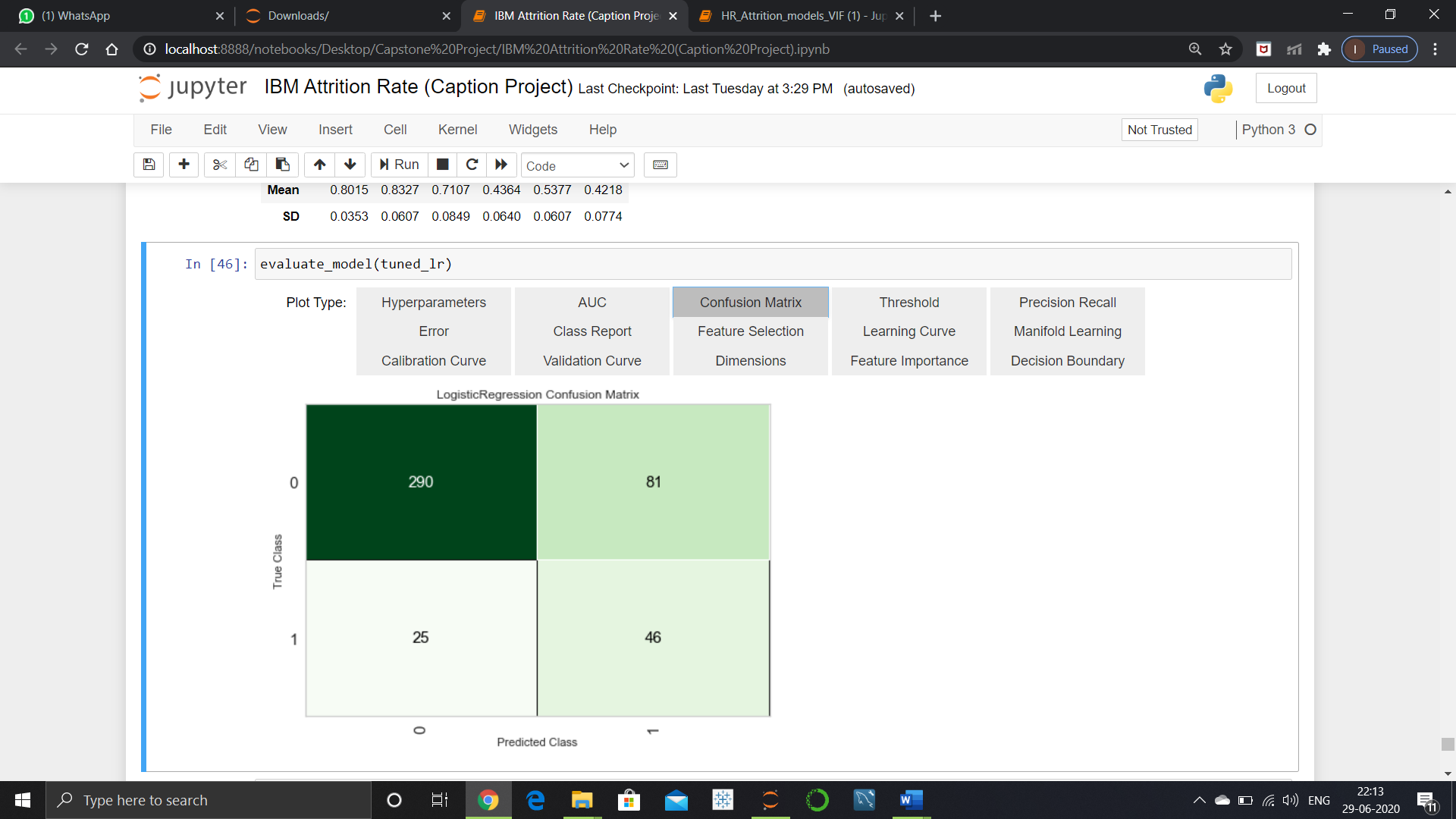
Alpha (False Positive) and Level of Significance are synonyms for Type 1 error and Type 2 error is commonly referred to as Beta (False Negative).

If we reject the null hypothesis when it is true, it refers to **Type 1** error.

Here, in this case let’s say, an employee does not wish to leave IBM but the model predicts him to potentially leave the company is an alpha error. This is basically a false alarm as the company need not worry but can leverage by proving better incentives or retention programs.

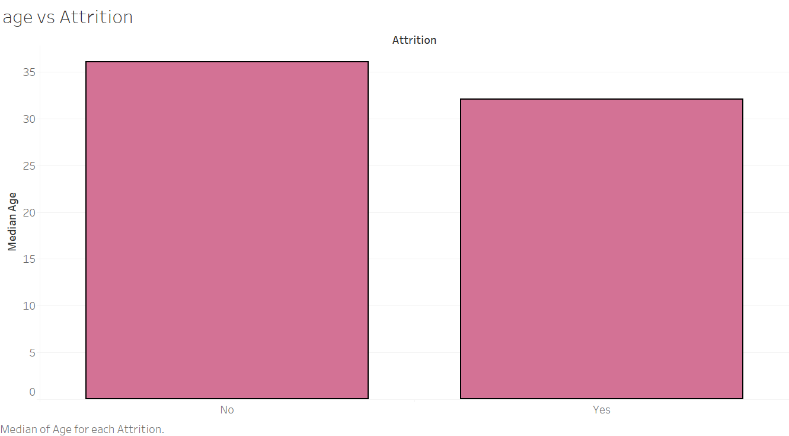
If we fail to reject the null hypothesis when it is false, it is a **Type 2** error.

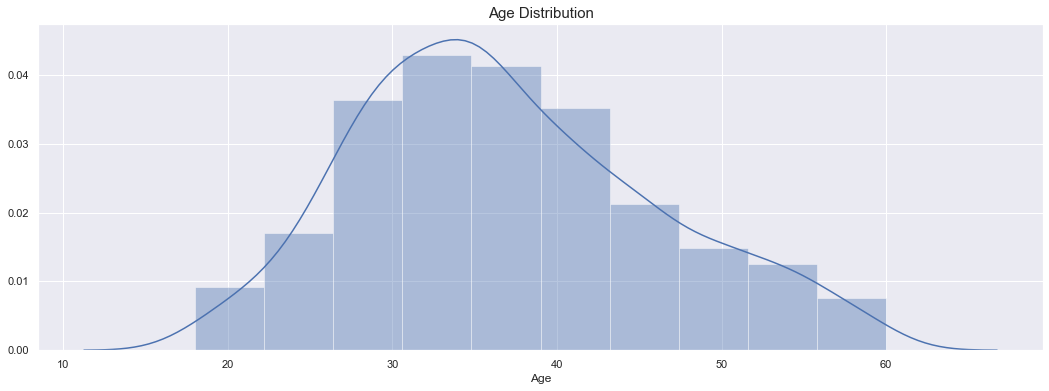
Here, an employee drafts his resignation mail and puts his papers down but the model wrongly predicts that he will not leave the company any time soon is a beta error. This is a missed opportunity as the company could have utilised the employee’s skills and also could have been prepared well in advan



**4.Exploratory Data Analysis**

**4.1 Data Exploration:**

**Age vs Attrition:**

Median age of attrition is approximately 32 years.

Attrition rate below age 20 is 58.82352941176471

Attrition rate of age between 20 and 30 is 5.1020408163265305

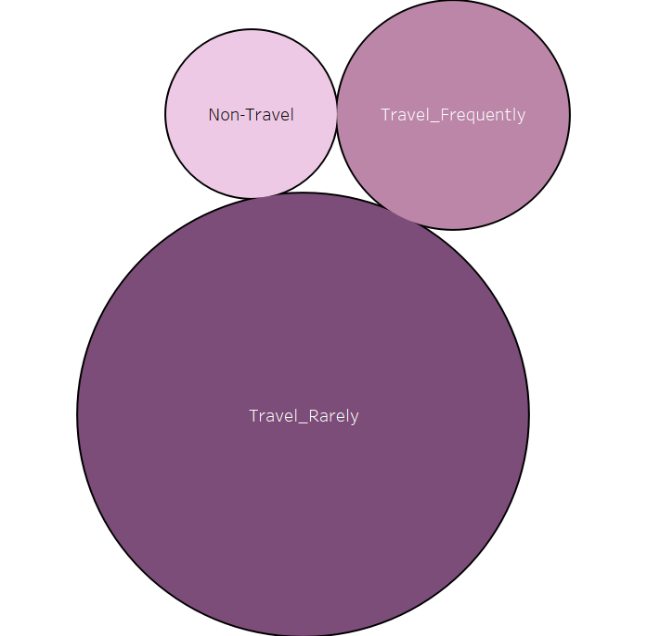
Attrition rate of age between 30 and 40 is 14.23487544483986

Attrition rate of age between 40 and 50 is 9.931506849315069

Attrition rate of age between 50 and 60 is 13.043478260869565

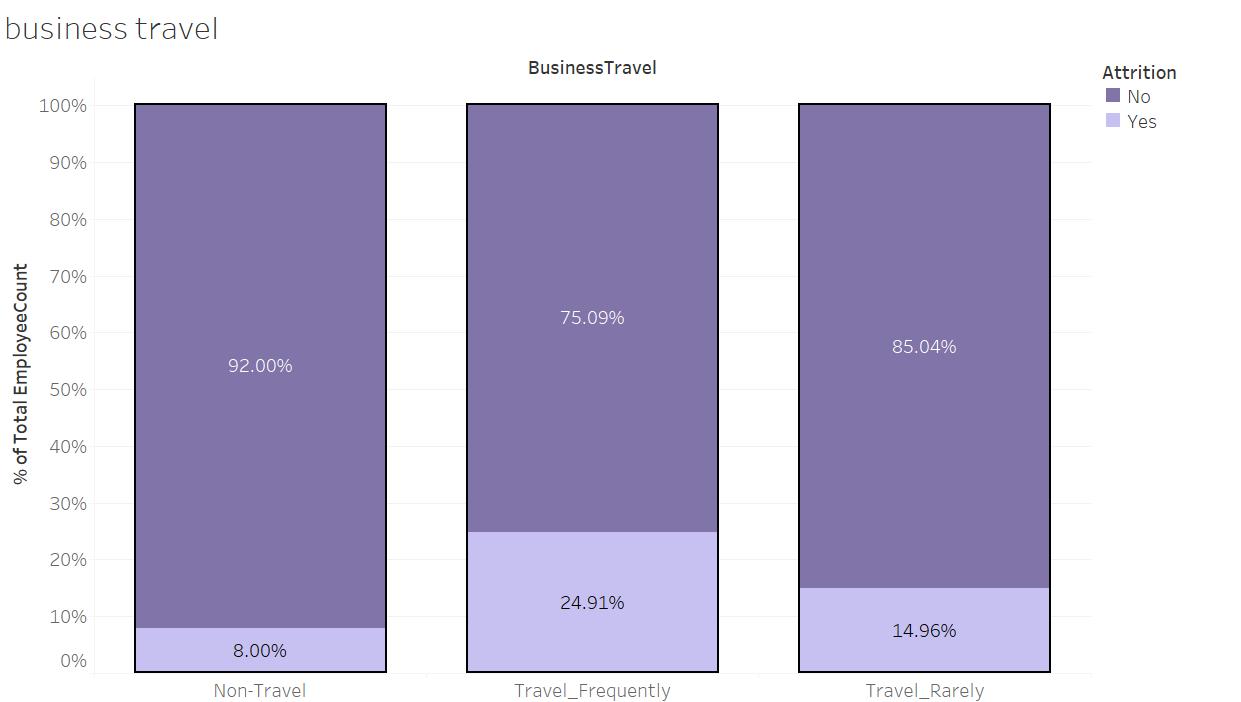
From the above statistics, we can see that the attrition is very high below the age of 20 and high in the age of 50 to 60.

**Business Travel vs Attrition:** This feature describes the type of travel the employee do. There are three categories in business travel.

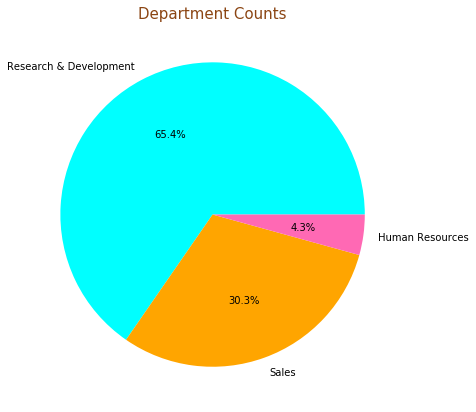
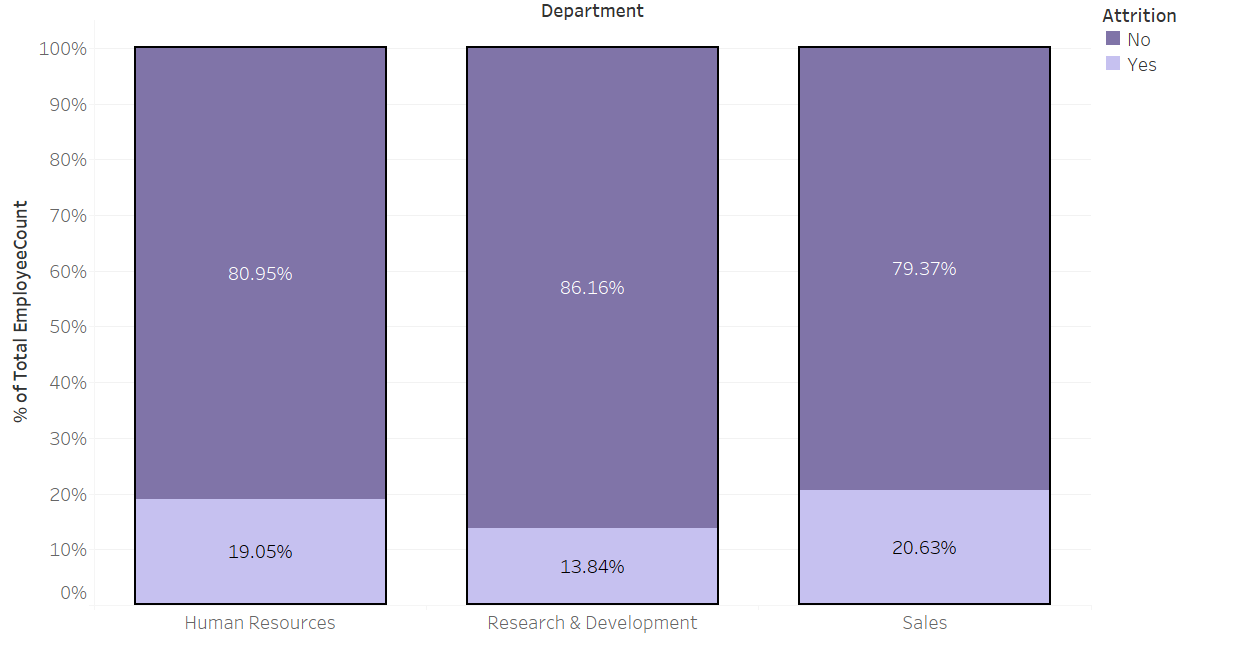
Travel Rarely 1043

Travel Frequently 277

Non-Travel 150



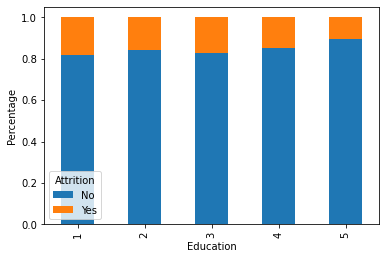
* Attrition rate when they travel rarely is 14.95%
* Attrition rate when they travel frequently is 24.9%
* Attrition rate when they don’t travel is 8%

**Department vs Attrition:** There are three departments namely Sales, Research Development, Human Resources.

* Attrition rate Sales is 20.62 %
* Attrition rate Research & Development is 13.83%
* Attrition rate Human Resources is 19.04
* **Distance from Home vs Attrition:** This feature depicts the total distance from home to office location. The lowest distance from their home to the workplace is 1 km and the farthest is 29kms. 75% of the people are within 14 kms distance from the workplace.



**Education vs Attrition:** There are five levels in education. They are:

1 – High school

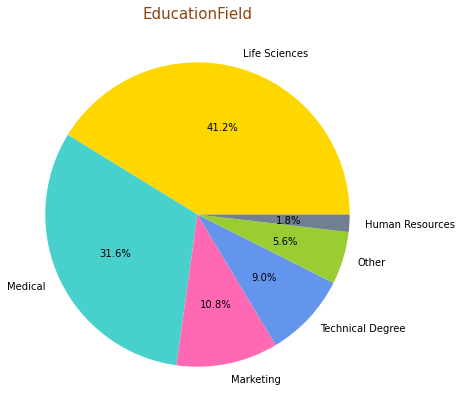
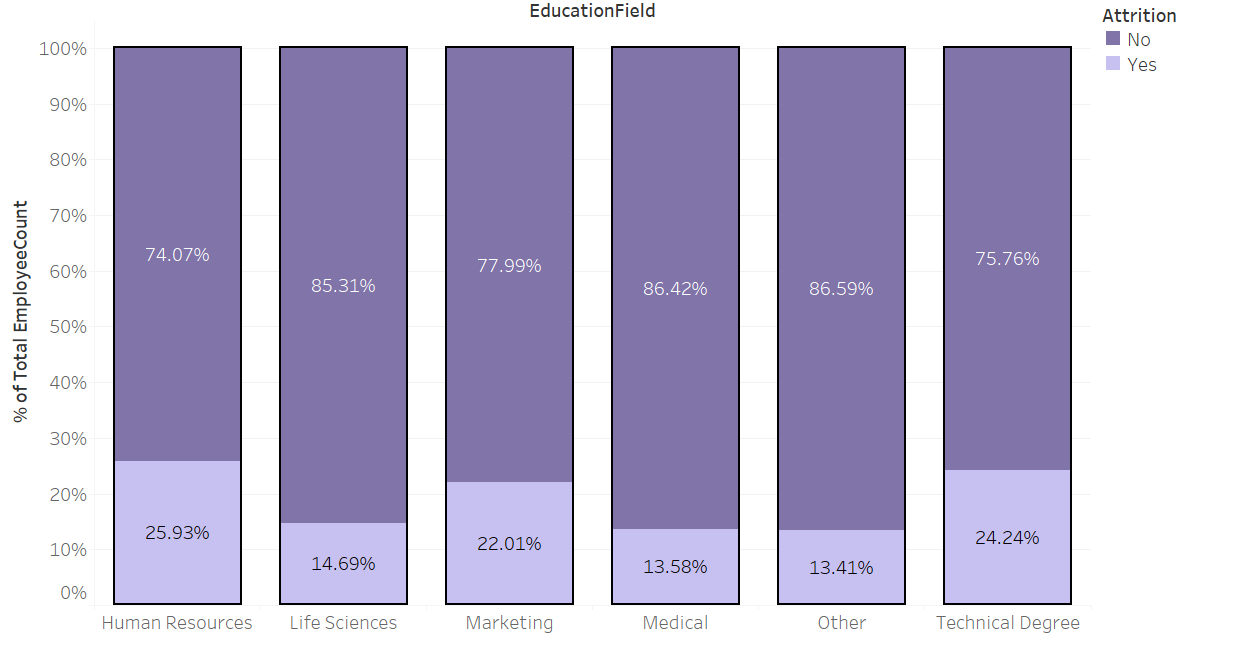
2 – Diploma

3 – Bachelor

4 – Masters

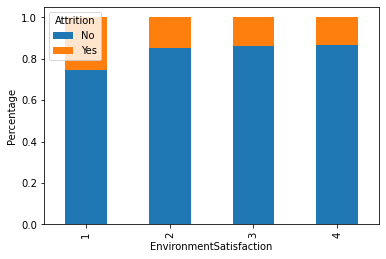
1. – Ph.D.

* Percentage of attrition rate w.r.t Education level 1 is 18%
* Percentage of attrition rate w.r.t Education level 2 is 16%
* Percentage of attrition rate w.r.t Education level 3 is 17%
* Percentage of attrition rate w.r.t Education level 4 is 15%
* Percentage of attrition rate w.r.t Education level 5 is 10%

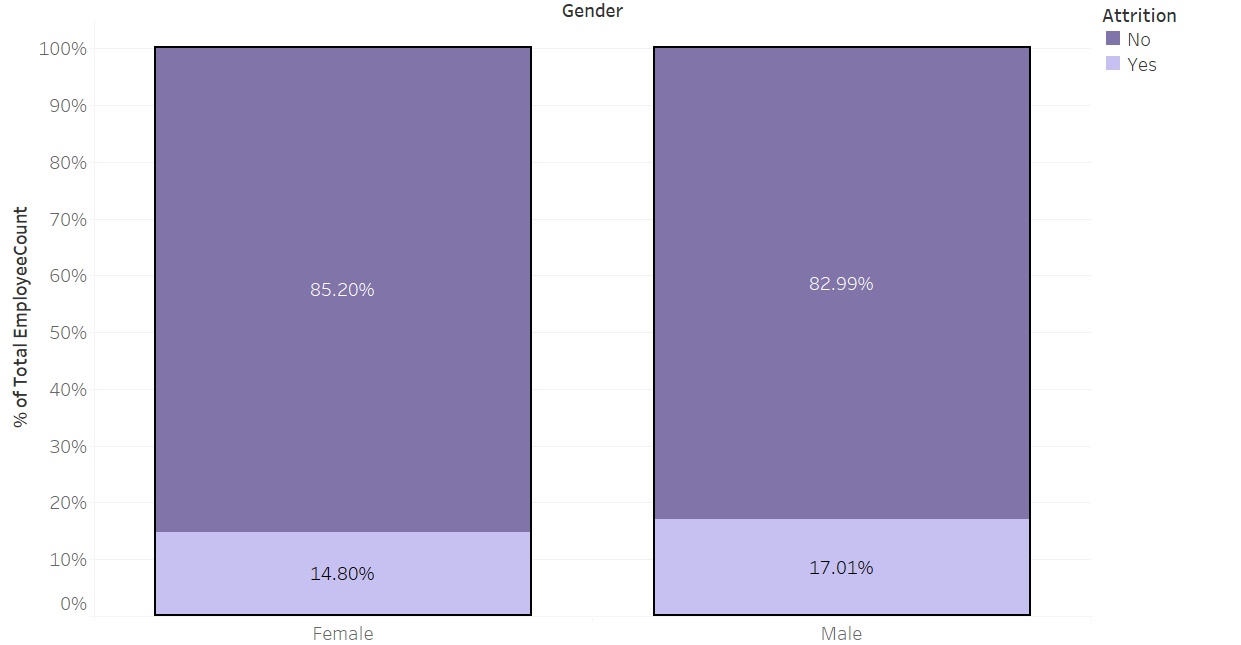
Level 3 is having highest count of people.

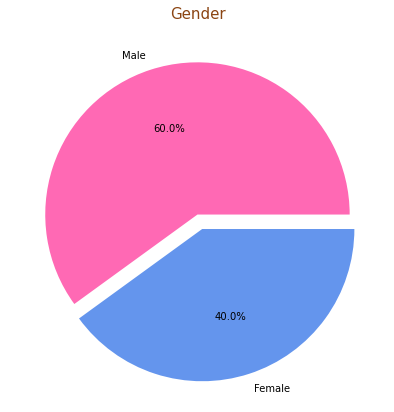
**Education Field vs Attrition:** There are 6 Education Fields.

* HR edu field attrition 25.93% (highest).
* Technical degree edu field attrition 24.24%.
* Marketing field attrition rate 22.01%
* Life science attrition rate 14.69%
* Medical field attrition rate 13.58%
* Other fields attrition rate 13.41%

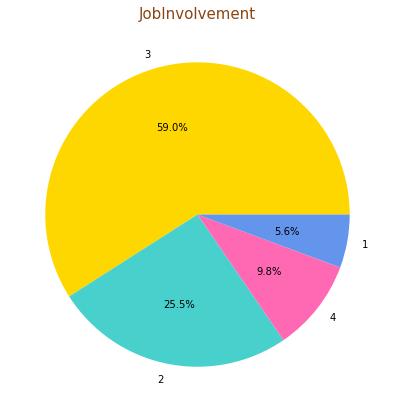
**Environment Satisfaction vs Attrition:** There are four levels in environment satisfaction. 1 being the lowest and 4 being the highest.

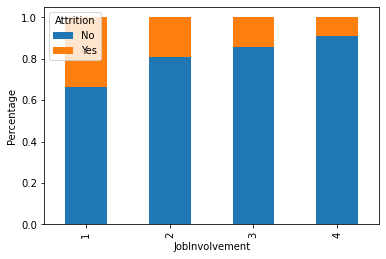
* Attrition for Environment Satisfaction level - 1 is 25.35%
* Attrition for Environment Satisfaction level - 2 is 14.98%
* Attrition for Environment Satisfaction level - 3 is 13.68%
* Attrition for Environment Satisfaction level - 4 is 13.45%

**Gender vs Attrition:** There are two genders male and female.

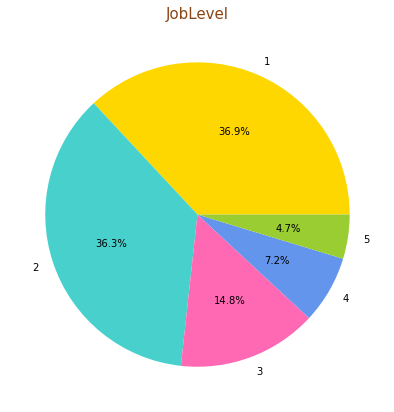
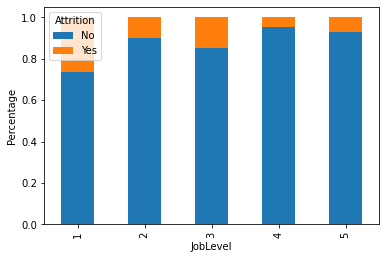


* Male attrition rate (17.01%) is higher than female (14.80%).

**Job Involvement vs Attrition:** This feature describes the rate at which the employee involved in their job. There are four levels 1 being the lowest and 4 being 4 the highest.

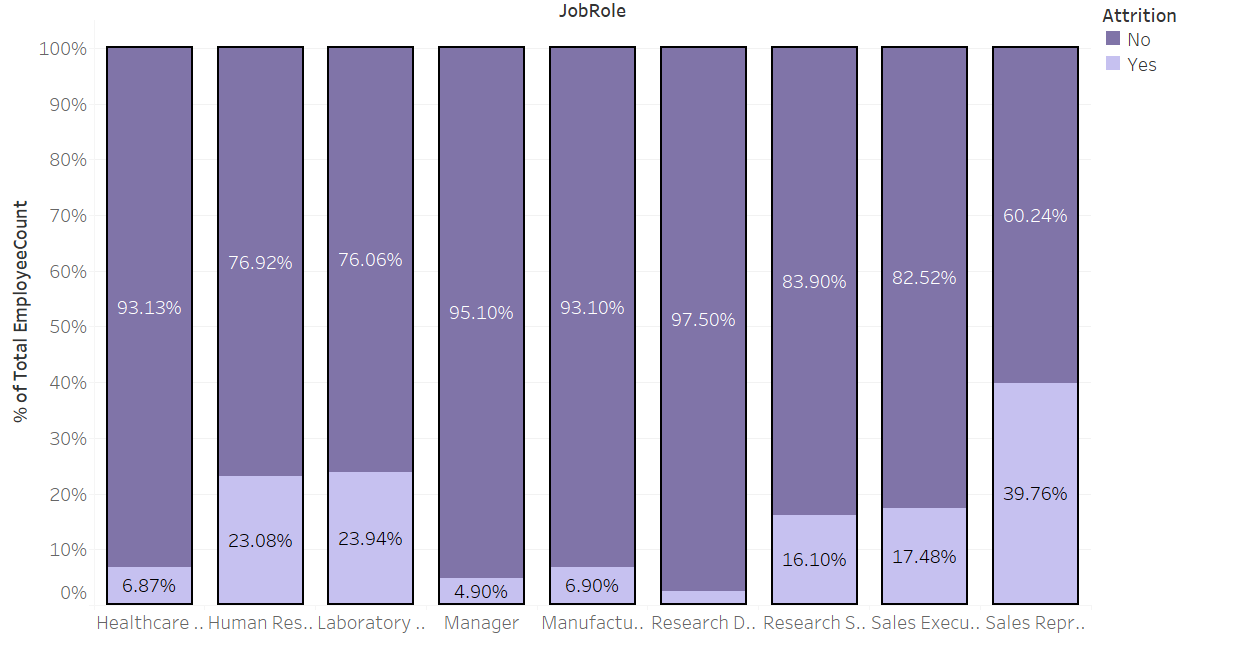
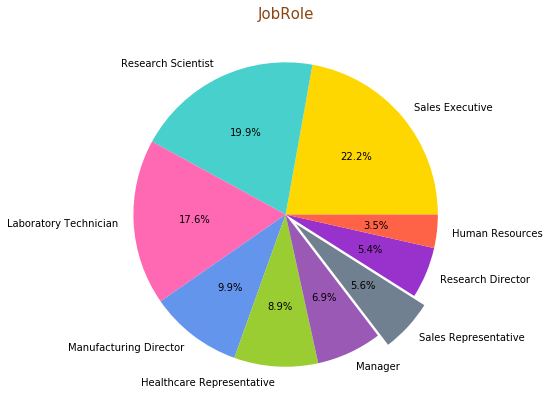


* Attrition rate for Job Involvement level 1 is 33.73%
* Attrition rate for Job Involvement level 2 is 18.93%
* Attrition rate for Job Involvement level 3 is 14.40%
* Attrition rate for Job Involvement level 4 is 9.02%

**Job Level vs Attrition:** There are 5 job levels, 1 being the lowest and 5 being the highest.

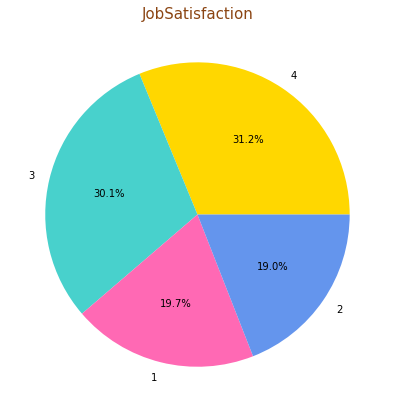
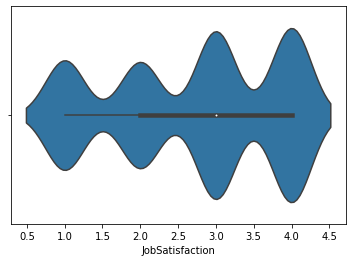
Most of the people are in job level 1 and level 2

* Attrition rate for Job level 1 is 26%
* Attrition rate for Job level 2 is 10%
* Attrition rate for Job level 3 is 15%
* Attrition rate for Job level 4 is 5%
* Attrition rate for Job level 5 is 7%

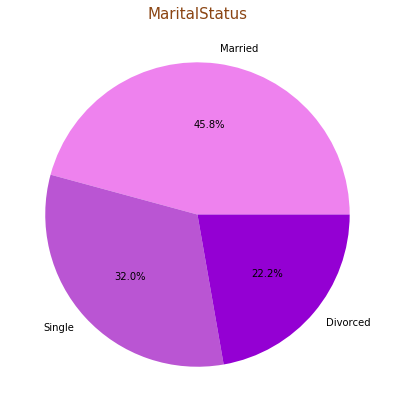
**Job Role vs Attrition:** There are 9 job roles.

* Attrition of Sales Executive: 17.48%
* Attrition of Research Scientist: 16.10 %
* Attrition of Laboratory Technician: 23.94%
* Attrition of Manufacturing Director: 6.90%
* Attrition of Healthcare Representative: 6.87%
* Attrition of Manager: 4.90%
* Attrition of Sales Representative: 39.76%
* Attrition of Research Director: 2.5%
* Attrition of Human Resources: 23.08 %

**Job Satisfaction vs Attrition:** The rate at which the employee is satisfied with the job. There are four levels 1 being the lowest and 4 being the highest.



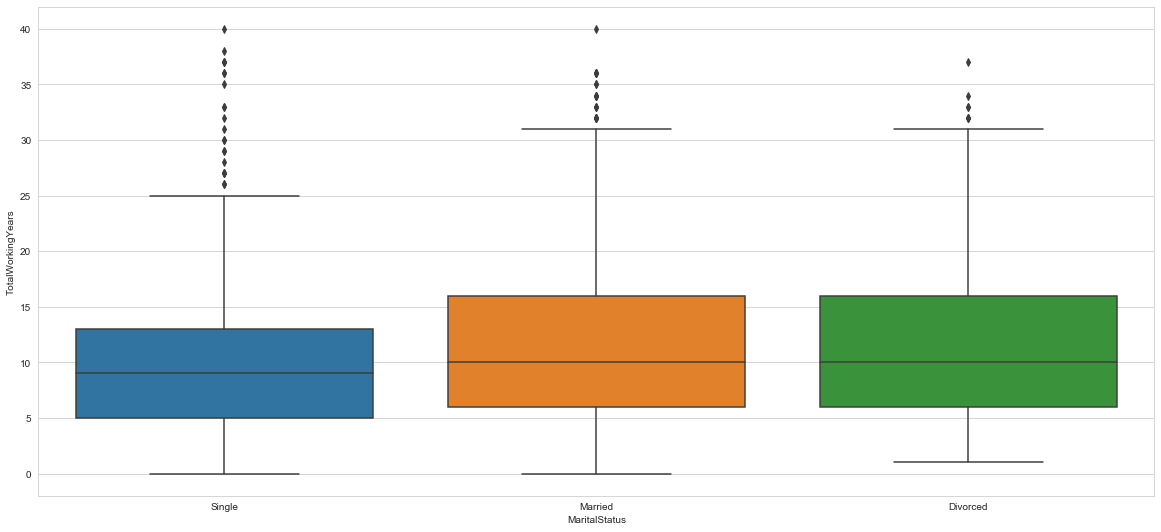
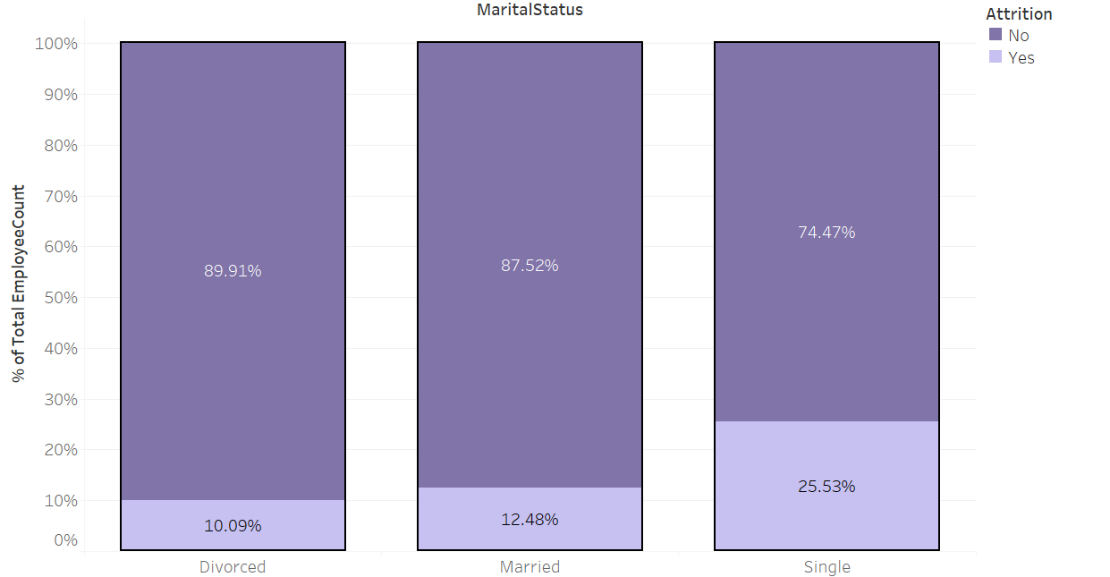
Median employees are with job satisfaction level 3.

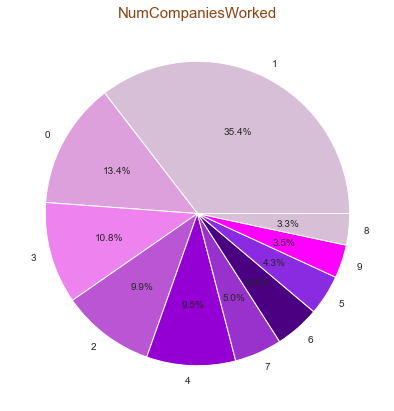
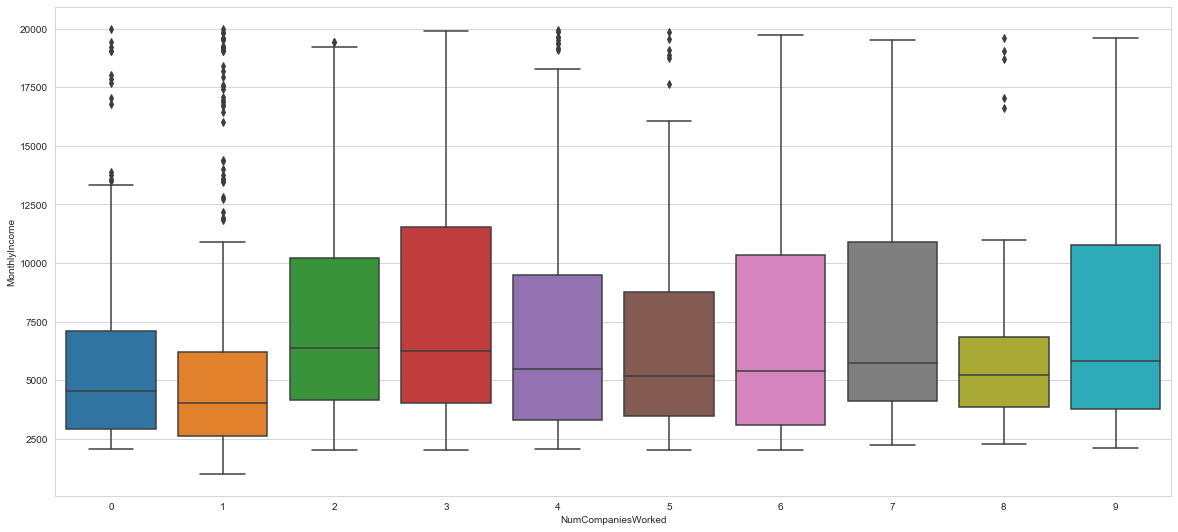
**Marital Status vs Attrition:** There are 3 categories in marital status.

They are:

1. Single
2. Married
3. Divorced

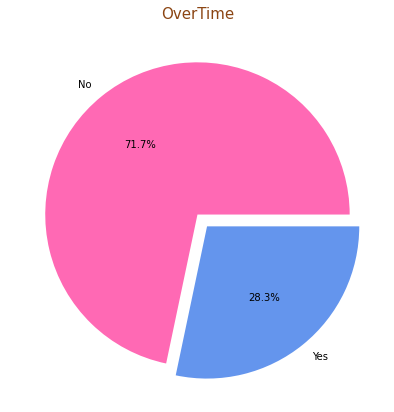
Single people are having less experience.

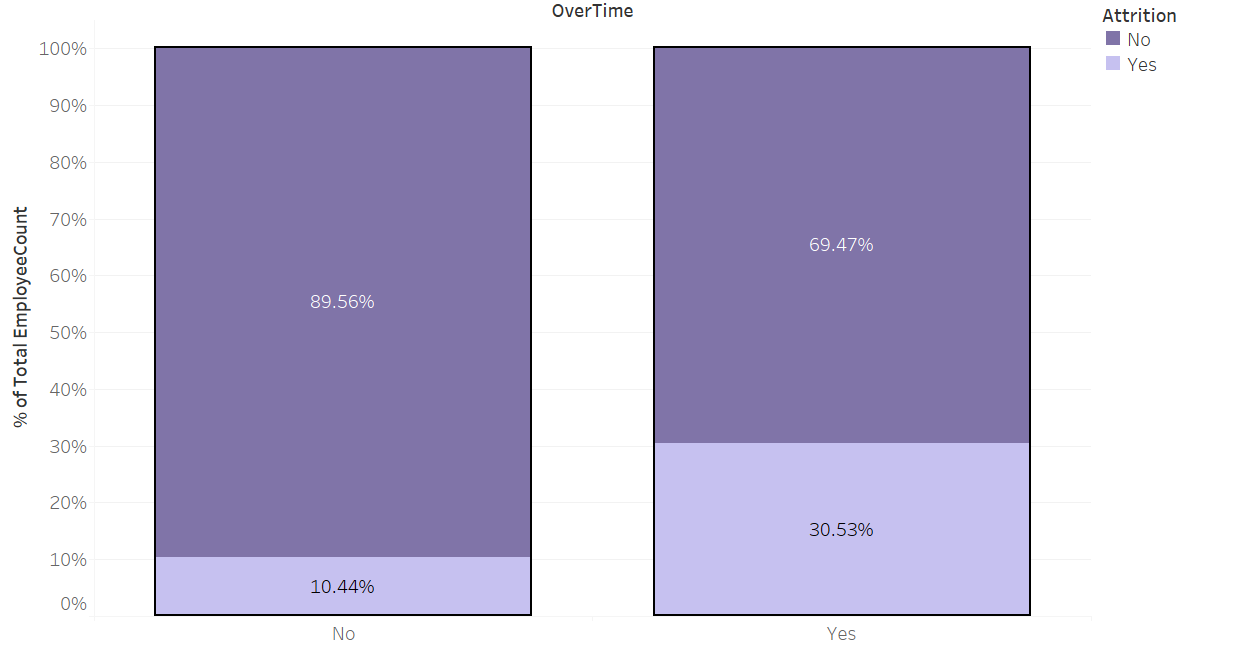
* Attrition rate when Marital Status Single is 25.53 %
* Attrition rate when Marital Status Married is 12.48%
* Attrition rate when Marital Status Divorced is 10.09%

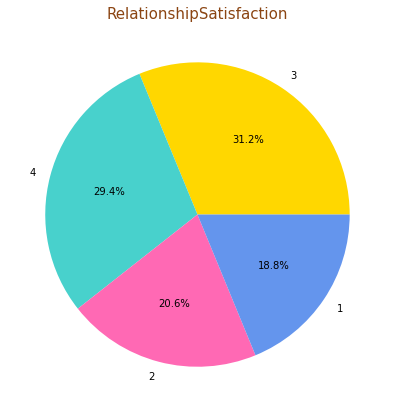
**Number of Companies Worked:** This field shows the number of companies that the employee worked before. The maximum limit is 9.

Some of the employees are having experience of working many companies though their salary is not so much increased.

**Overtime:** This is a Boolean feature with 'Yes' and 'No'.

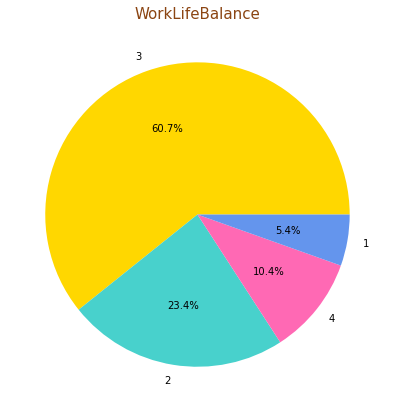
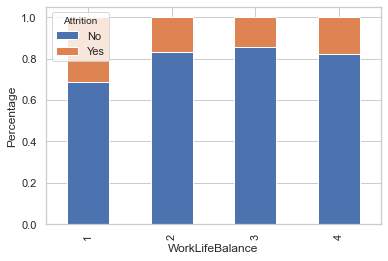
* Attrition rate if they have done Overtime is 30.53%
* Attrition rate if they have not done Overtime is 10.44%



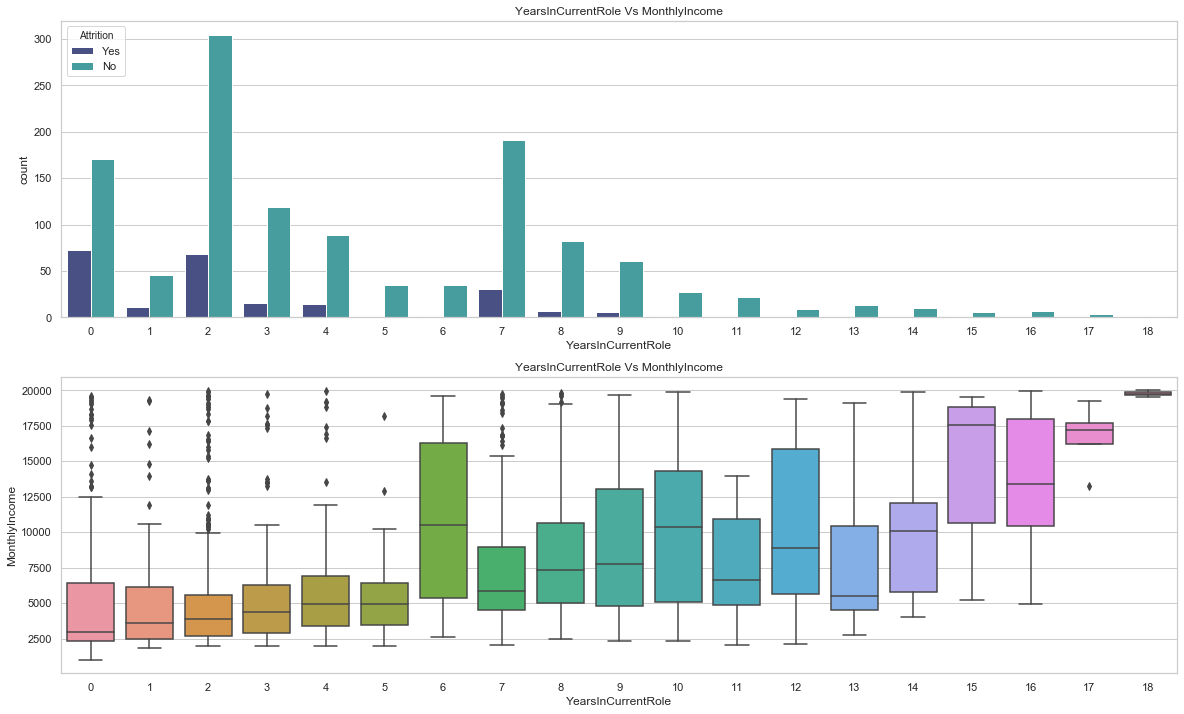
**Relationship Satisfaction:** This field depicts the level of satisfaction of the employees in their relationships. 1 being lowest and 4 being the highest.



* Attrition rate when Relationship Satisfaction 1 is 21%
* Attrition rate when Relationship Satisfaction 2 is 15%
* Attrition rate when Relationship Satisfaction 3 is 15%
* Attrition rate when Relationship Satisfaction 4 is 15%

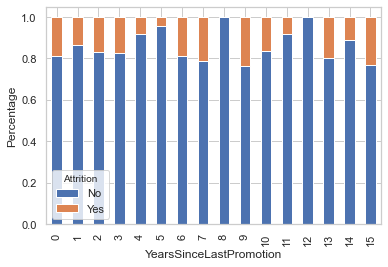
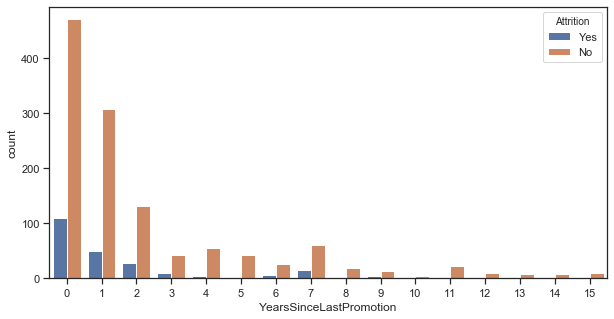
**Worked Life Balance vs Attrition:** This field describes the rate at which the employee balances work and life. There are four levels 1 being lowest and 4 being the highest.

* Attrition rate when Work life balance is of level 1 is 31.25%
* Attrition rate when Work life balance is of level 2 is 16.86%
* Attrition rate when Work life balance is of 3 is 14.22%
* Attrition rate when Work life balance is of 4 is 17.64%

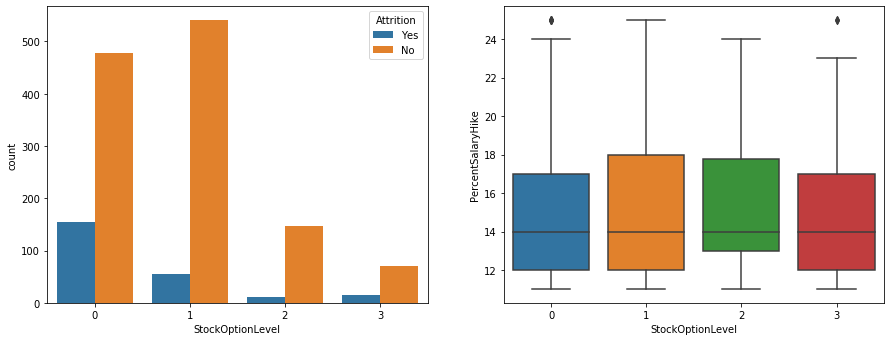
**Years at Current Role vs Attrition:** This field describes the number of years the employee is working in the current role.

* Attrition rate for working in current role for less than 5 years is 20.06%
* Attrition rate for working in current role between 5 to 10 years is 11.08 %
* Attrition rate for working in current role between 10 and 15 years is 5.26%

Employee with less ‘years with current role’ salary is also low and their attrition is high.

**Years Since Last Promotion vs Attrition:** This feature describes the number of years since the last promotion. In the dataset, the years range from 0 to 18.

* Attrition rate for 0 years since last promotion is 19 %
* Attrition rate for 1 years since last promotion is 14 %
* Attrition rate for 2 years since last promotion is 17 %
* Attrition rate for 3 years since last promotion is 17 %
* Attrition rate for 4 years since last promotion is 8 %
* Attrition rate for 5 years since last promotion is 4%
* Attrition rate for 6 years since last promotion is 19%
* Attrition rate for 7 years since last promotion is 21%
* Attrition rate for 9 years since last promotion is 24%
* Attrition rate for 10 years since last promotion is 17 %
* Attrition rate for 11 years since last promotion is 8 %
* Attrition rate for 13 years since last promotion is 20%
* Attrition rate for 14 years since last promotion is 11 %
* Attrition rate for 15 years since last promotion is 23 %
* **Stock Option Level vs Attrition:**

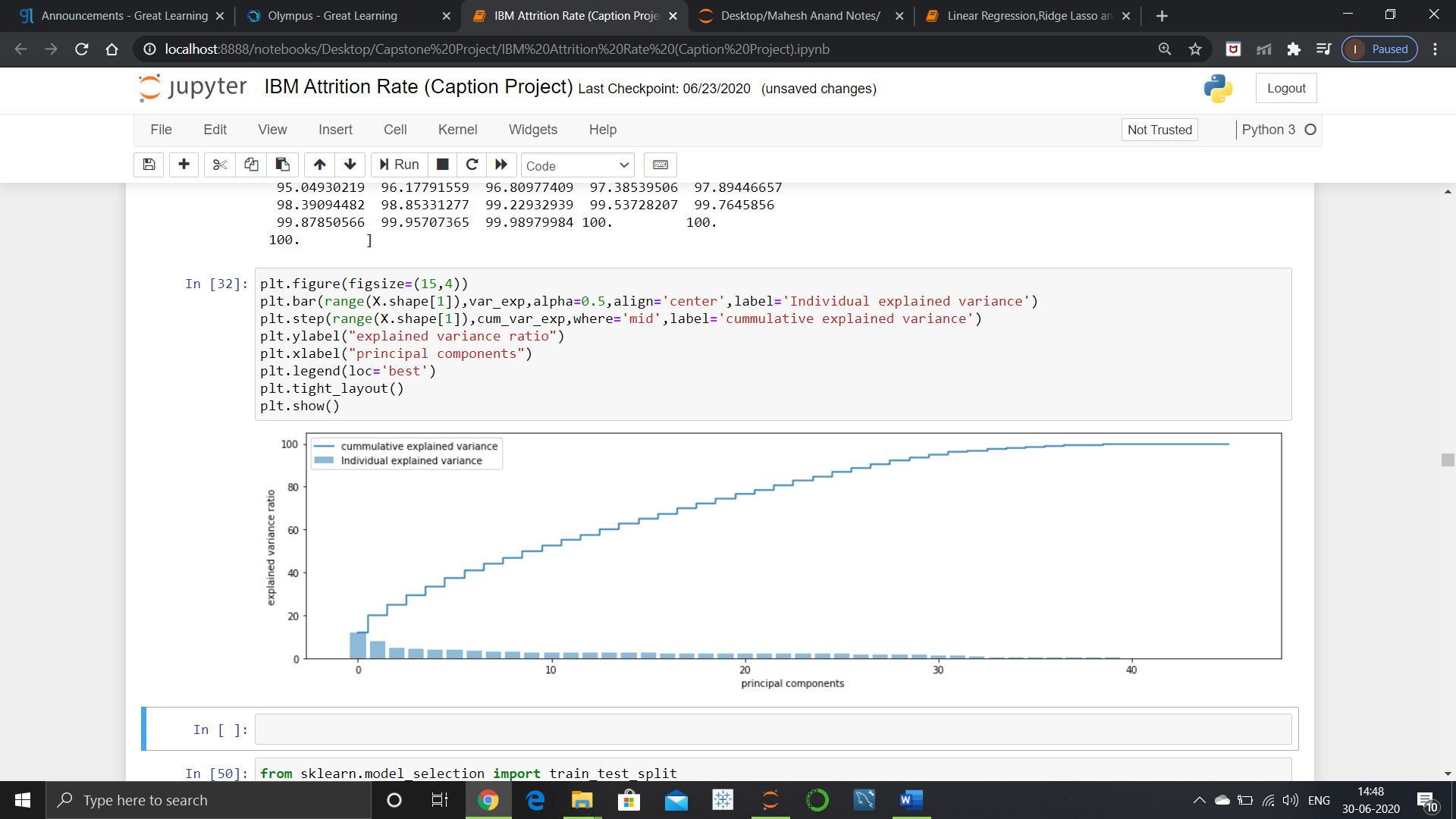
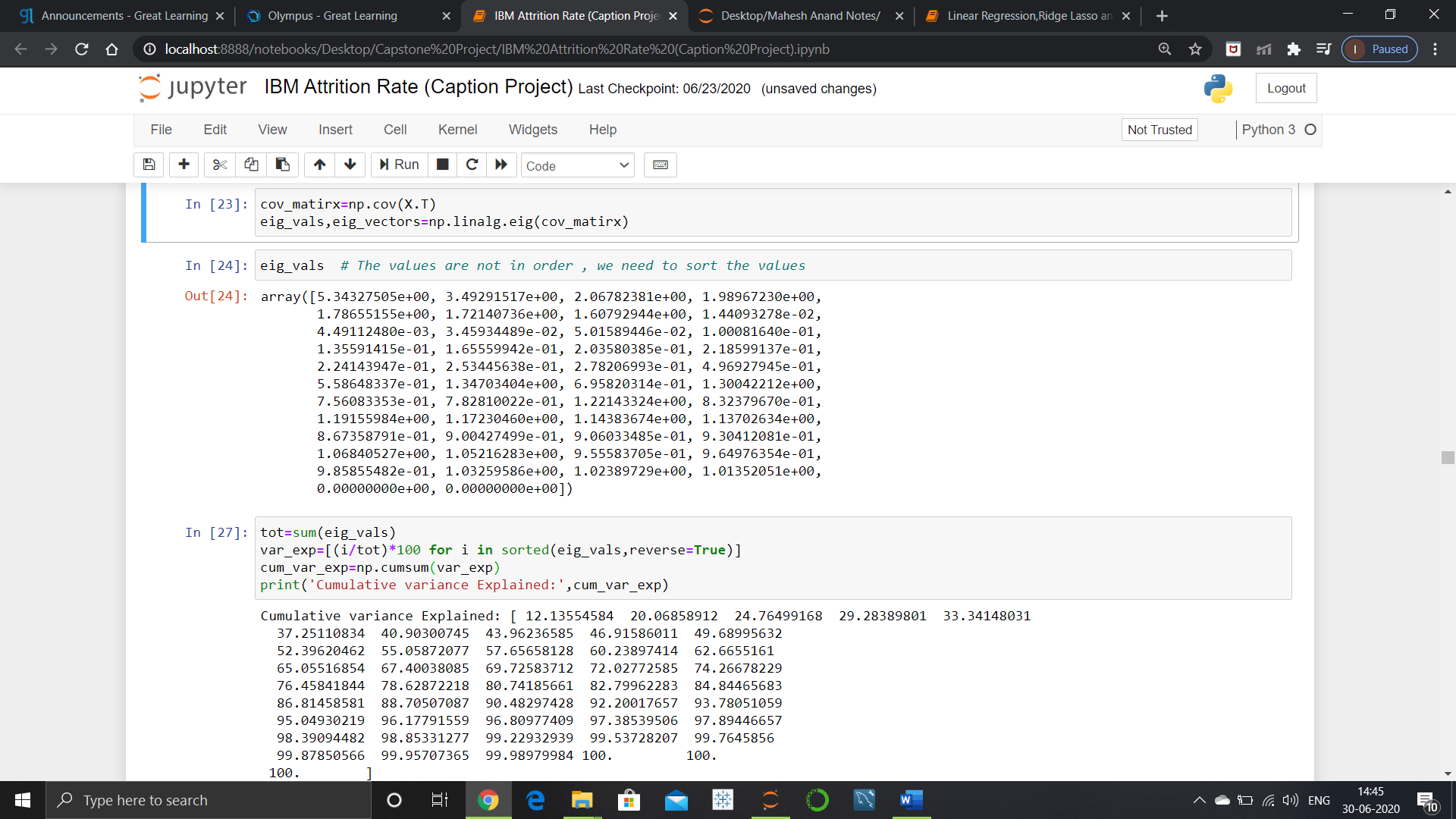


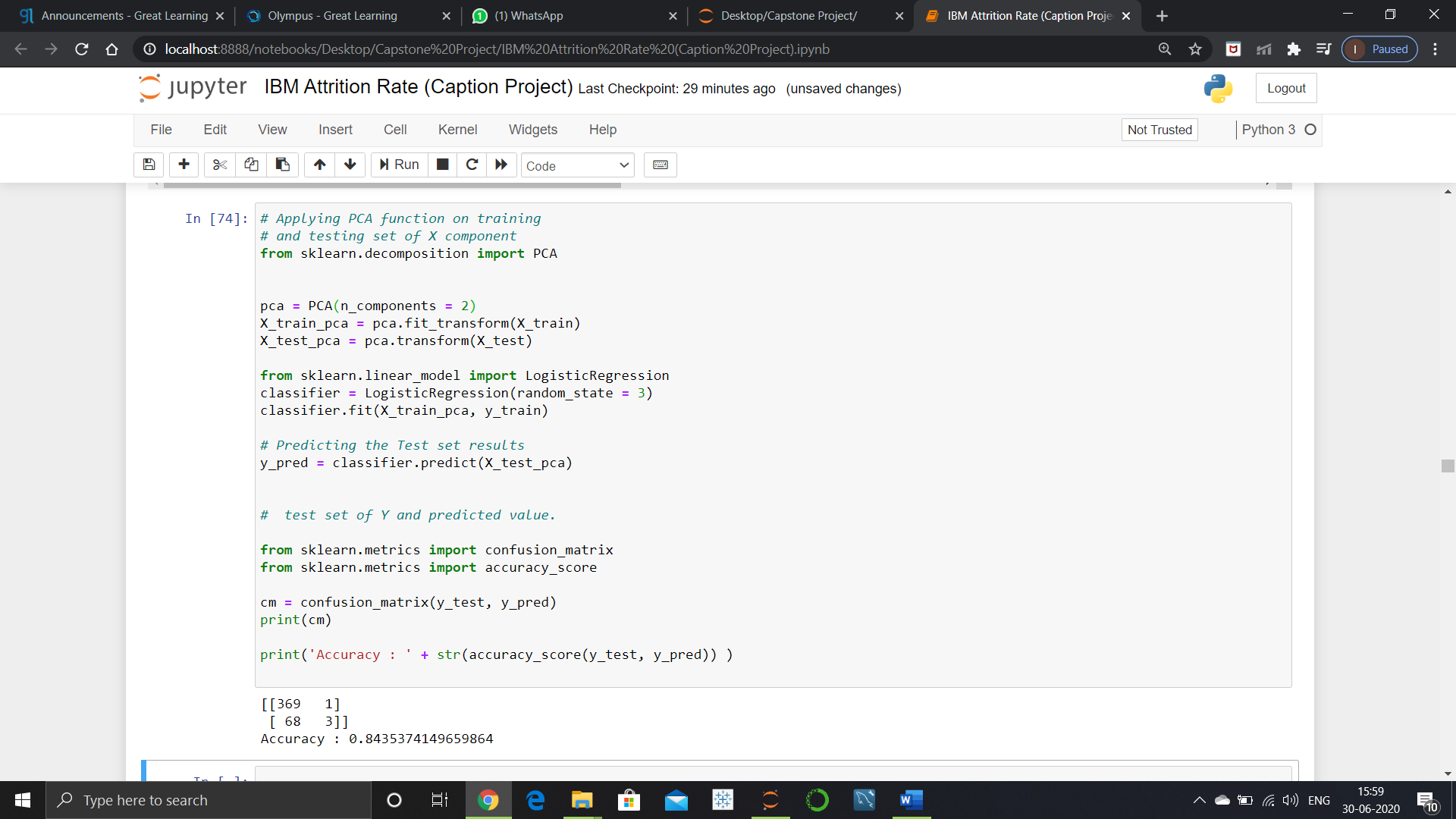
* Attrition Rate of Employees with StockOptionLevel 0: 24.40%
* Attrition Rate of Employees with StockOptionLevel 1: 9.39%
* Attrition Rate of Employees with StockOptionLevel 2: 7.59%
* Attrition Rate of Employees with StockOptionLevel 3: 17.64%

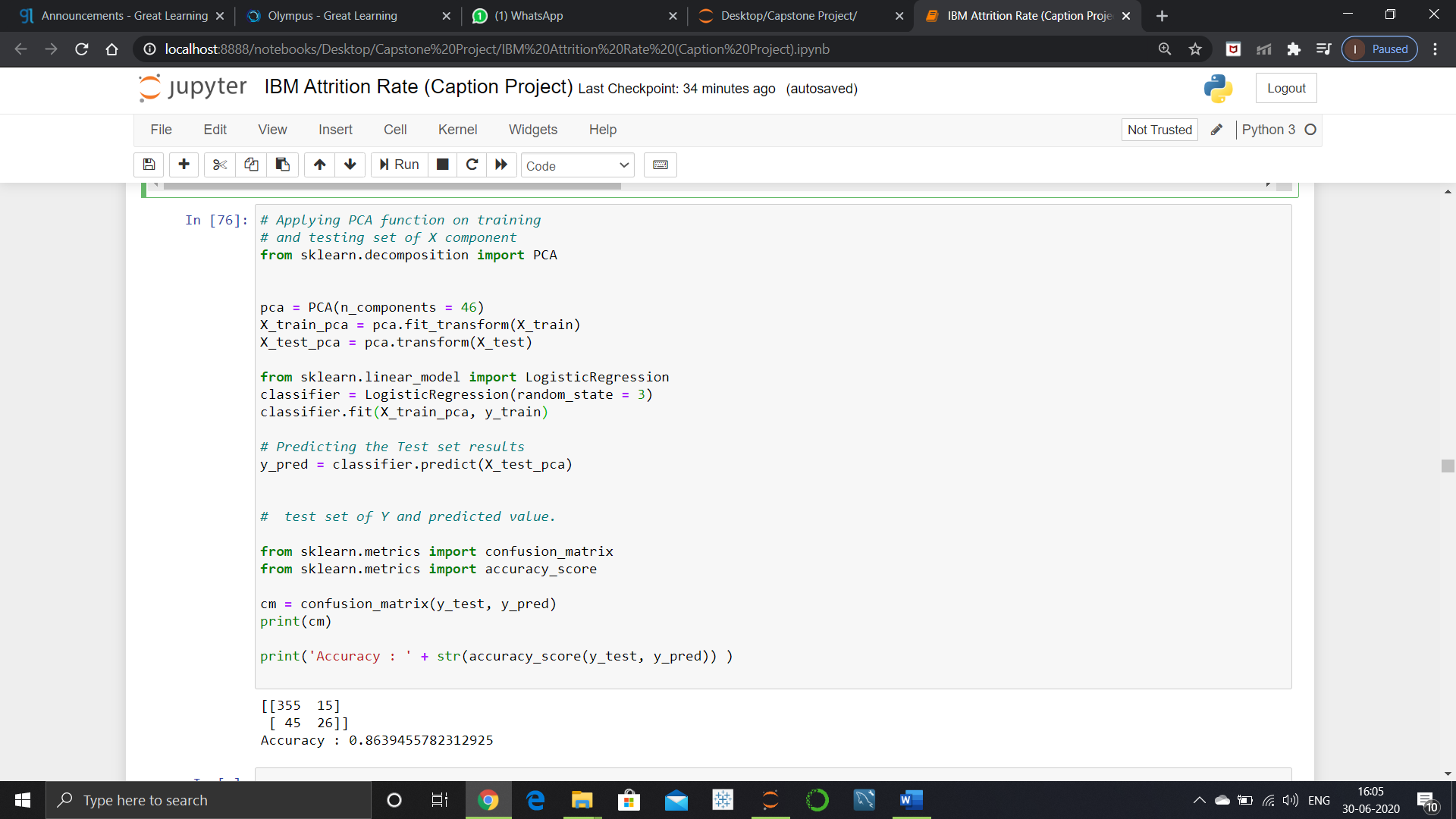
Employees with Stock Option as 0, also have a high tendency of

leaving the organisation.

**4.2 Principal Component Analysis (PCA)**

The main objective is to counter multicollinearity effect which usually occurs when the data set is large, thereby helping the model with additional information, PCA helps reduce redundant data by acting as a controlling mechanism and due to this dimensionality reduction is a by-product the process. PCA tries to capture the unique information which is present in ± 2 or ± 3 sigma. Scaling the data set is absolutely crucial as well .

PCA DOES NOT DROP ANY RECORDS OR ATTRIBUTES !!

Here, we can notice that the first 37 columns out of 46 carry 99% of the information. We can therefore choose to drop the last 9 columns and therefore not only removing multicollinearity but achieving dimensionality reduction too.

**A mere 2 % increase of overall accuracy when n componenets increased from 2 to 46**

**5 Feature Engineering**

**5.1 Outliers**

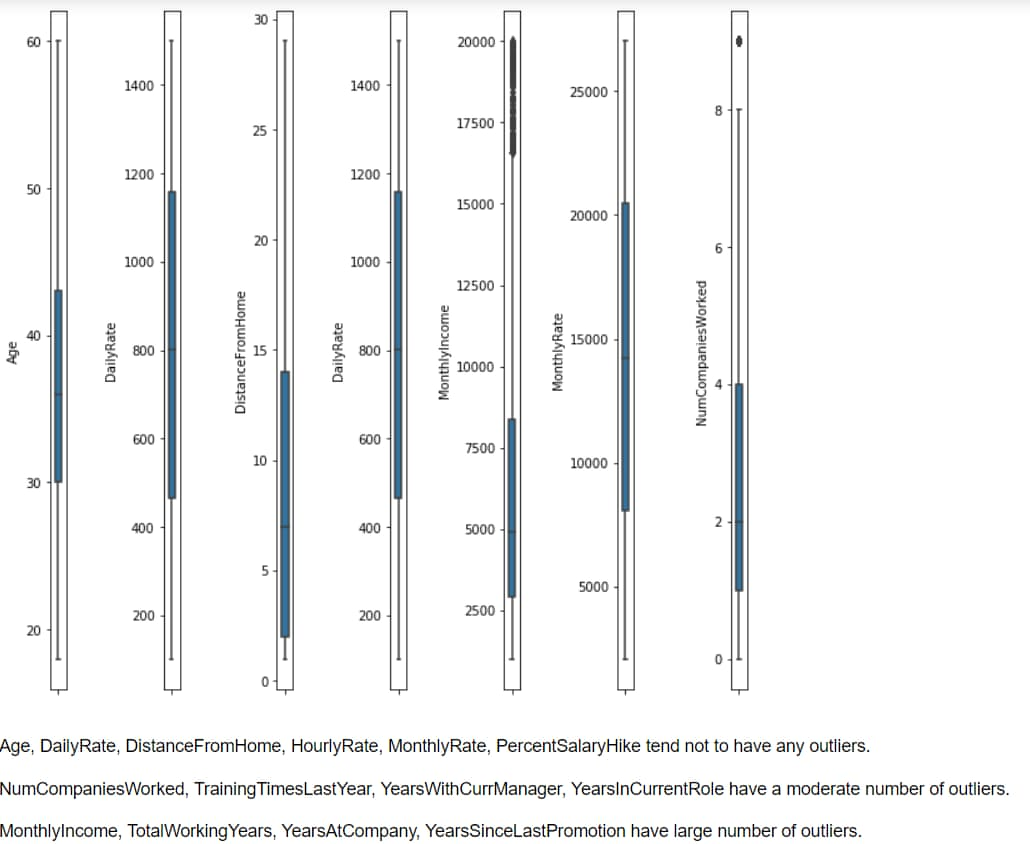
An outlier is a data point in a data set that is distant from all other observation.

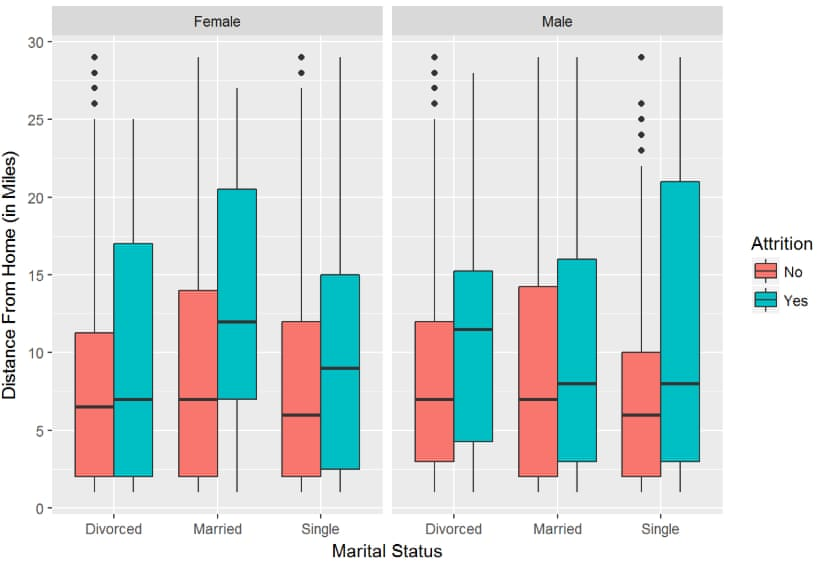
In statistics, an **outlier** is an observation point that is distant from other observations.

## **Discover outliers with visualization tools**

1. Box-plot

In descriptive statistics, a **box plot** is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have **lines extending vertically** from the boxes (whiskers) **indicating** **variability** outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. **Outliers** may be **plotted** as **individual** points.





This graph is combined of 3 features which includes gender, martial status and distance from home .

This collection of boxplots shed light on how distance from home could positively correlate with higher attrition level. We can see the median of attired employees had to commute longer than those who stayed on each instance for both gender alongside their marital status.

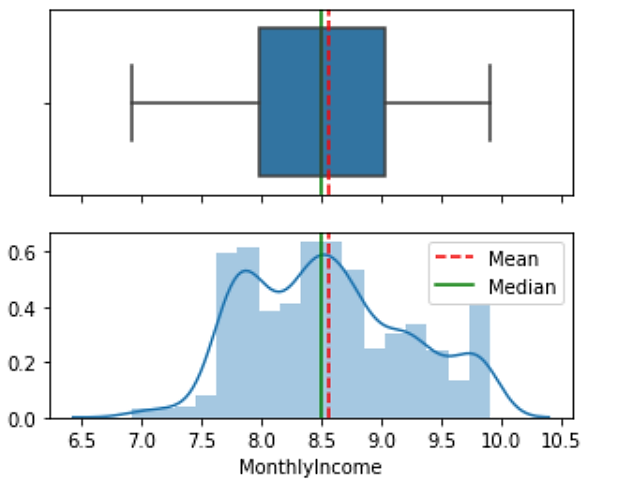
**5.2 Data Transformations**

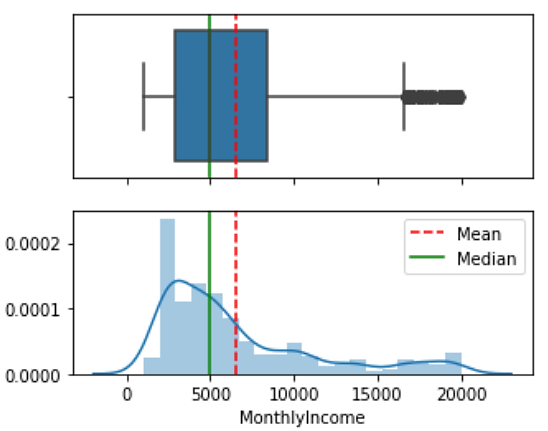
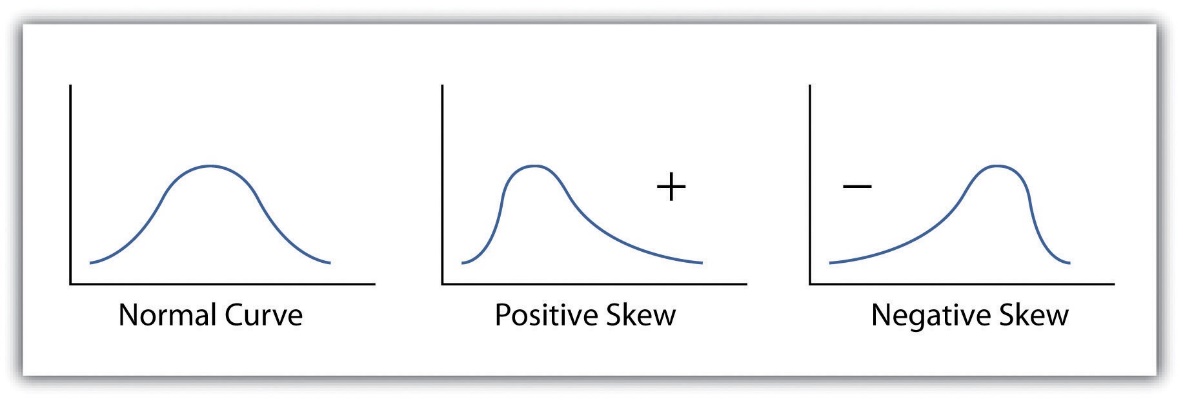
Data Transformation is the process of converting the data from one form of structure to another form using mathematical functions. Examples of some transformers used are log transform, power transform, square root transform, box-cox transform etc.

Let’s understand why do we need transformation.

Whenever we have outliers, we get skewed distribution. It becomes very important to identify the reason for the presence of outliers or skewness. Let’s consider a variable “Monthly Income” from our Attrition dataset. From the graphs below, we see that monthly income is right skewed as mean is affected by the presence higher outliers. Reason for these values being outliers could be that these values might have been wrongly entered, therefore we could drop the entries or we can use any of the imputations correct the distribution. Another thought could be monthly income is not same for all employees and it depends one’s skill-set, job level, experience etc. Therefore, in a company most people get median income, and higher pay is given to employees with skills, experience etc. As we build the model with the presence of these extreme values are not captured. To make are model capture this useful information we need our data distribution to follow normality. This is the situation where we use data transformation techniques to make the distribution normal and not losing the useful information.

Before data transformation After data transformation

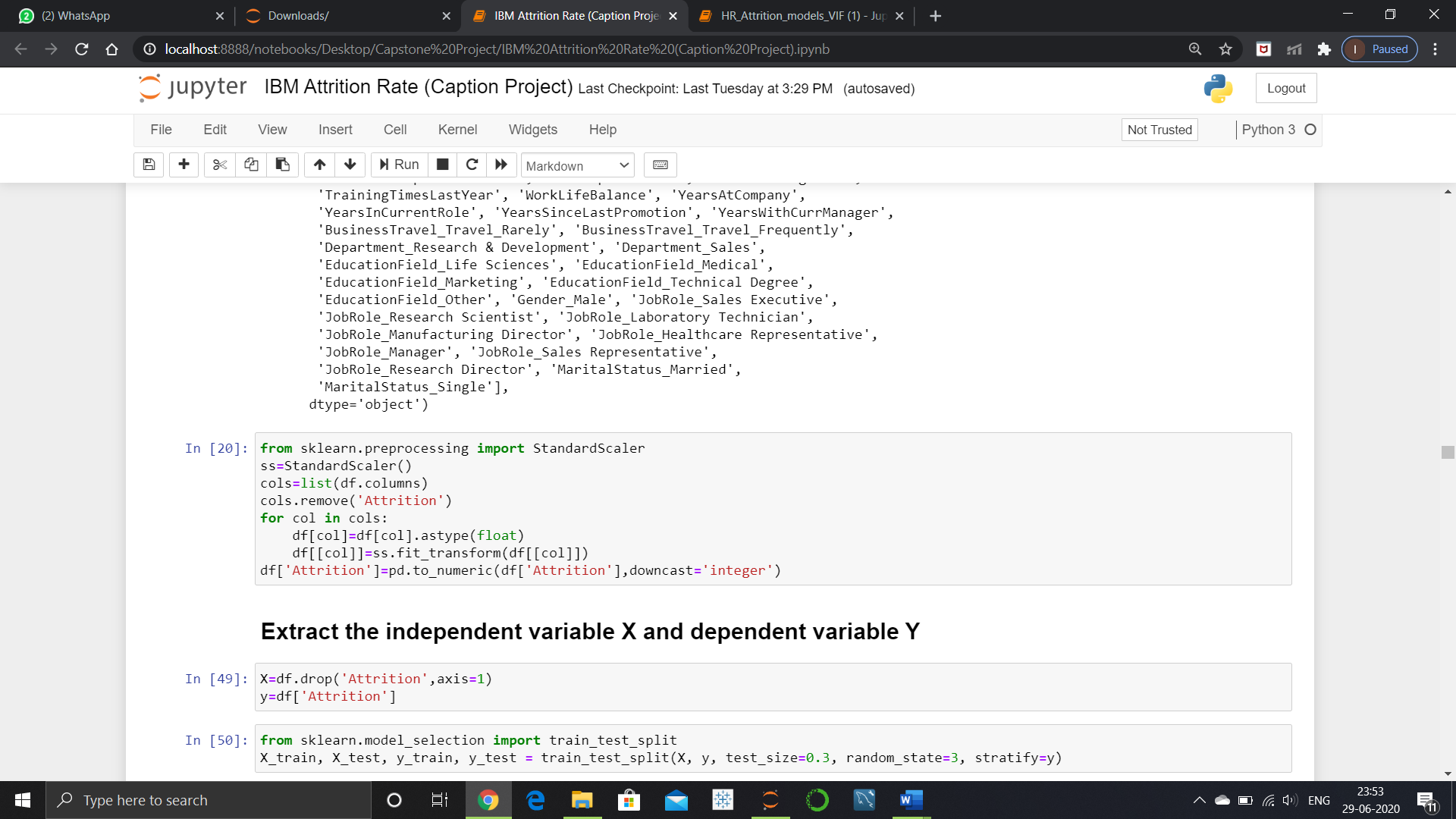
Skewness: 1.36 Skewness: 0.28

Graph after data transformation has reduced the skewness from 1.36 to 0.28, which a considerable drop. This data can be further used to build the model.

For positive (right) skewed distribution we can use log transform, square-root transform, inverse transform. For negative(left) skewed various power transforms such as square, cube etc, box-cox transform can be applied.

**5.3 Scaling the Data**

Scaling the data is a very important pre-processing step used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre- processing step. There are 3 types of SCALING that we generally prefer: Standard Scaling, Min-Max Scaling

and Robust Scaling. In our data, we have used Standard Scaling.

**3.2.1 Standard Scaling:**

This is one of the most commonly used types of scaling that we use in machine learning. It basically works on “Z-score” which means that Standard Scaler removes the mean and scales the data to unit variance. The mean of the data will be “0” and standard deviation would be equal to 1.

**5.4 Feature Extraction:**

**Feature Extraction** is the process of constructing new features from existing data to train a machine learning model. Having and engineering good features will allow us to most accurately represent the underlying structure of the data and therefore create the best model.

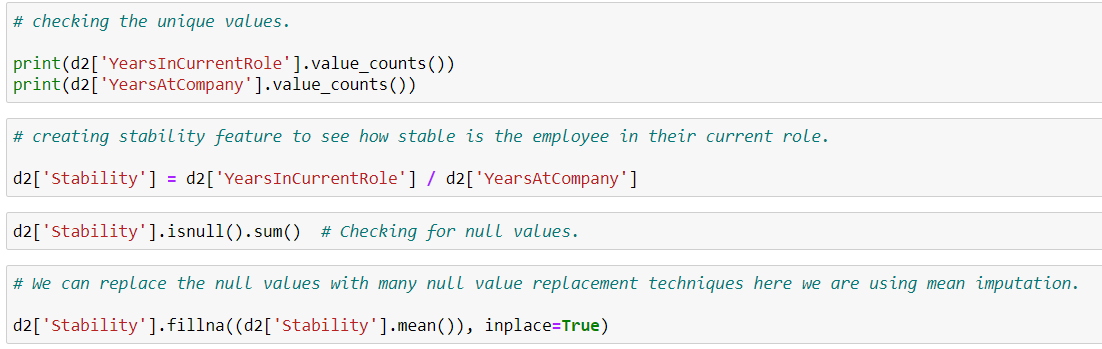
Features can be engineered by decomposing or splitting features, from external data sources, or aggregating or combining features to create new features.

**1.AllSatisfaction\_mean**

After looking at the distinct values of all 5-predictor feature (RelationshipSatisfaction, EnvironmentSatisfaction, JobSatisfaction, JobInvolvement, WorkLifeBalance) we see same pattern being followed so we took average of all the feature and made a new feature ‘AllSatisfaction\_mean’.

**2.Stability**

We wanted to get an idea of how the employee’s stability is affecting the attrition.

****Stability here refers to an employee’s average years in current role from the time they joined the company. This will also give us idea of how employee’s promotion is affecting the attrition rate as well.

**3.Inc/dist\_unit**

****With this new feature we wanted to have a general idea of how the attrition rate is affected by the salary an employee to the distance he travels for work. We will get the trend of the employee attrition with respect to income per unit distance.

**5.5 Feature Selection**

Feature Selection is one of the main steps of the pre-processing phase as the features which we choose directly affects the model performance. While some of the features seem to be less useful in terms of the context, others seem to equally useful. The better features we use the better our model will perform. Irrelevant or partially relevant features can negatively impact model performance. Feature selection and Data cleaning should be the first and most important step of your model designing. Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

**5.5.1 Benefits of Feature selection:**

Reduces Overfitting:Less redundant data means less opportunity to make

decisions based on noise.

Improves Accuracy**:**Less misleading data means modelling accuracy improves.

Reduces Training Time**:**Fewer data points reduce algorithm complexity and

algorithms train faster.

**Feature selection technique that we used:**

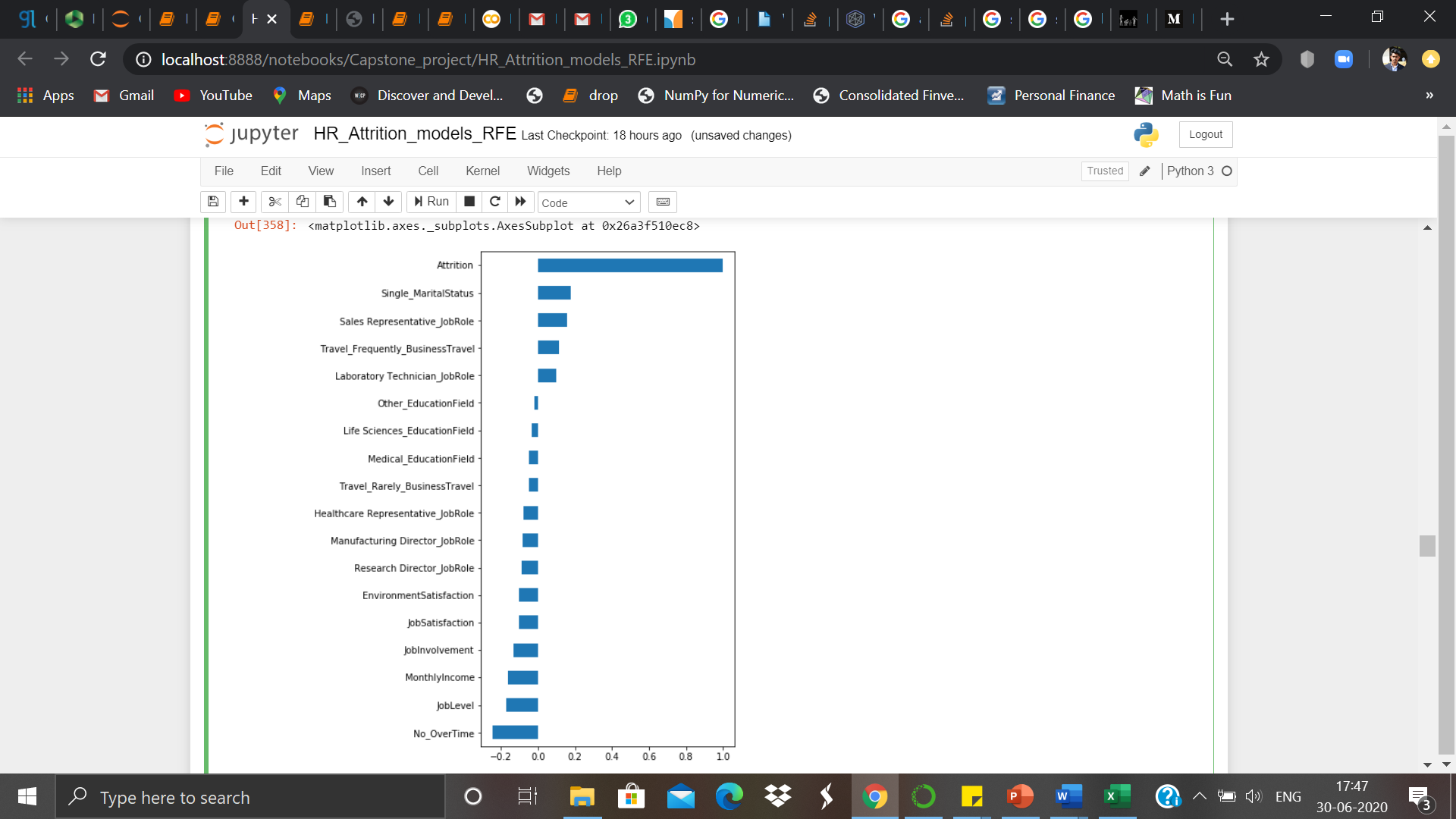
**5.5.2 Recursive Feature Elimination (RFE):**

* Recursive Feature Elimination (RFE) takes as input the instance of a machine learning model and the final desired number of features to use. It then recursively reduces the number of features to use by ranking them using the machine learning model accuracy as metrics.
* Creating a for loop in which the number of input features is our variable, it could then be possible to find out the optimal number of features our model needs by keeping track of the accuracy registered in each loop iteration.
* Using RFE support method, we can then find out the names of the features which have been evaluated as most important (rfe.support return a boolean list in which TRUE represent that a feature is considered as important and FALSE represent that a feature is not considered important.

**5.5.2.1 Selecting most important features using RFE:**

|  |  |
| --- | --- |
| 1 | Age |
| 2 | Distance From Home |
| 3 | Environment Satisfaction |
| 4 | Job Involvement |
| 5 | Job Satisfaction |
| 6 | Num Companies Worked |
| 7 | Years In Current Role |
| 8 | Years Since Last Promotion |
| 9 | Years With Curr Manager |
| 10 | Business Travel-Travel Frequently |
| 11 | Business Travel-Travel Rarely |
| 12 | Job Role-Laboratory Technician |
| 13 | Job Role-Sales Executive |
| 14 | Job Role-Sales Representative |
| 15 | Marital Status-Single |
| 16 | Overtime-Yes |

These are the important features we get after applying RFE.



6.1 **Model Implementation (GAME PLAN)**



This phase corresponds to building and execution of models. At the beginning, a process flow was planned and followed. Below is the chart showing the process flow.

In the beginning we have explored the overview of the data in the EDA phase, which has made us get the best out of the data.

Following methods were used to carry out the exploration:

**Statistical Analysis**: To check the significance of a feature with respect to the target we perform different statistical analysis techniques by checking the probability of null hypothesis being true. (that the feature has no influence on the target variable).

For categorical variables, chi-square test has been performed while for numerical variable One-way Anova was used to check the significance of the model. Below are the variables that passed the statistical tests.

|  |  |  |  |
| --- | --- | --- | --- |
| Attrition | MaritalStatus | JobInvolvement | WorkLifeBalance |
| BusinessTravel | OverTime | JobSatisfaction | YearsAtCompany |
| Department | Age | MonthlyIncome | YearsInCurrentRole |
| EducationField | DailyRate | StockOptionLevel | YearsWithCurrManager |
| JobLevel | DistanceFromHome | TotalWorkingYears |  |
| JobRole | EnvironmentSatisfaction | TrainingTimesLastYear |  |

Once selected, nominal categorical variables were converted to numerical variables using panda’s get\_dummies, and ordinal categories were Label Encoded. A total of 36 columns were formed after encoding.

**Feature Selection**:

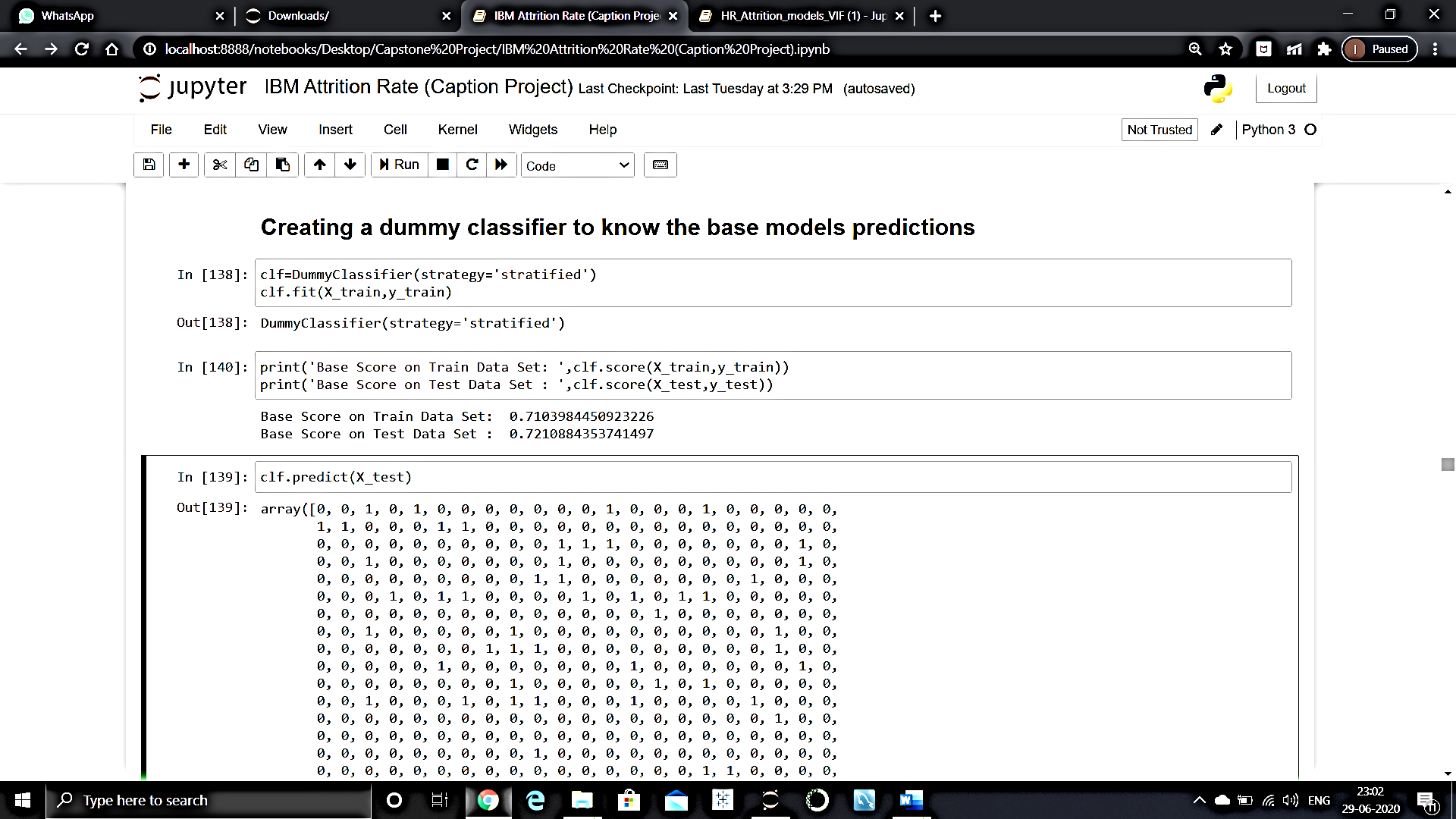
**Correlation:**

A data-set with good attributes should not have a correlation among themselves. Correlation matrix was used to find the highly correlated attributes, as these variables can adversely affect the models as there is a possibility of them carrying the same information (redundant information). Therefore, it’s important to handle them. After checking the heat map (Visual representation of correlation) and Variance Inflation Factor (VIF depicts the correlation features with each other), it was observed that there were many features highly correlated with each other. Under the condition of them carrying the same information, these attributes were eliminated.

|  |  |  |
| --- | --- | --- |
| Age | StockOptionLevel | YearsWithCurrManager |
| DailyRate | TotalWorkingYears | Travel\_Rarely\_BusinessTravel |
| DistanceFromHome | TrainingTimesLastYear | Travel\_Frequently\_BusinessTravel |
| EnvironmentSatisfaction | WorkLifeBalance | Sales\_Department |
| JobInvolvement | YearsAtCompany | Medical\_EducationField |
| JobSatisfaction | Technical Degree\_EducationField | Manager\_JobRole |
| YearsInCurrentRole | Other\_EducationField | Sales Representative\_JobRole |
| Marketing\_EducationField | Research Scientist\_JobRole | Research Director\_JobRole |
| Healthcare Representative\_JobRole | Laboratory Technician\_JobRole | Married\_MaritalStatus |
| No\_OverTime | Manufacturing Director\_JobRole | Single\_MaritalStatus |

**Class Imbalance**:

Identified the class imbalance of the data set. It was observed that there is approximately 84% of observations that belong to ’No’ class in dataset. Hence the model will be biased while learning. To overcome this, we can either have under sampling or over sampling technique. Since, our data set is not large we go for oversampling technique SMOTE was used to overcome the class Imbalance. For oversampling, SMOTE uses the k-nearest neighbours’ value on the sample of dataset. The value of k=5 was observed to be the best value to oversample the minority class in this data.



**K-fold Cross Validation**:

As this dataset has limited observations, there was a high chance of overfitting in the data. To eliminate this, k-fold repeated cross validation was used. With this technique, the sample of data was randomly divided into equally sized samples by the algorithm, giving out a single best sample for the testing.

**Hyperparameter Tuning**:

The hyperparameter tuning was done by implementing Randomized Grid Search with cross validation to get the best fit tuned hyperparameters of the classification models. These parameters are usually hard-coded in the models, and vary depending on the model used.

After the exploration, distributed data was divided into training and testing data sets. A dummy model (using dummy classifier) was modelled to set a threshold score, wherein a model built should not yield a score less than this score. Our train accuracy score for the dummy model was 0.71, and that of the test score is 0.72

**6.Model Building:**

Linear models require feature selection in order to not perform brilliantly.

**6.2.1 Logistic Regression**

**6.2.1.1 ASSUMPTIONS OF LOGISTIC REGRESSION:**

Logistic regression is quite different than linear regression in that it does not make several of the key assumptions that linear and general linear models (as well as other ordinary least squares algorithm based models) hold so close.

* Logistic regression does not require a linear relationship between the dependent and independent variables.
* The error terms (residuals) do not need to be normally distributed.
* Homoscedasticity is not required.
* The dependent variable in logistic regression is not measured on an interval or ratio scale.

However, logistic regression still shares some assumptions with linear regression, and some additions of its own.

1.ASSUMPTION OF APPROPRIATE OUTCOME STRUCTURE

Binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.

2.ASSUMPTION OF OBSERVATION INDEPENDENCE

Logistic regression requires the observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data.

3.ASSUMPTION OF THE ABSENCE OF MULTICOLLINEARITY

Logistic regression requires there to be little or no multi-collinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other.

4.ASSUMPTION OF LINEARITY OF INDEPENDENT VARIABLES AND LOG ODDS

Logistic regression assumes linearity of independent variables and log odds. Although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.

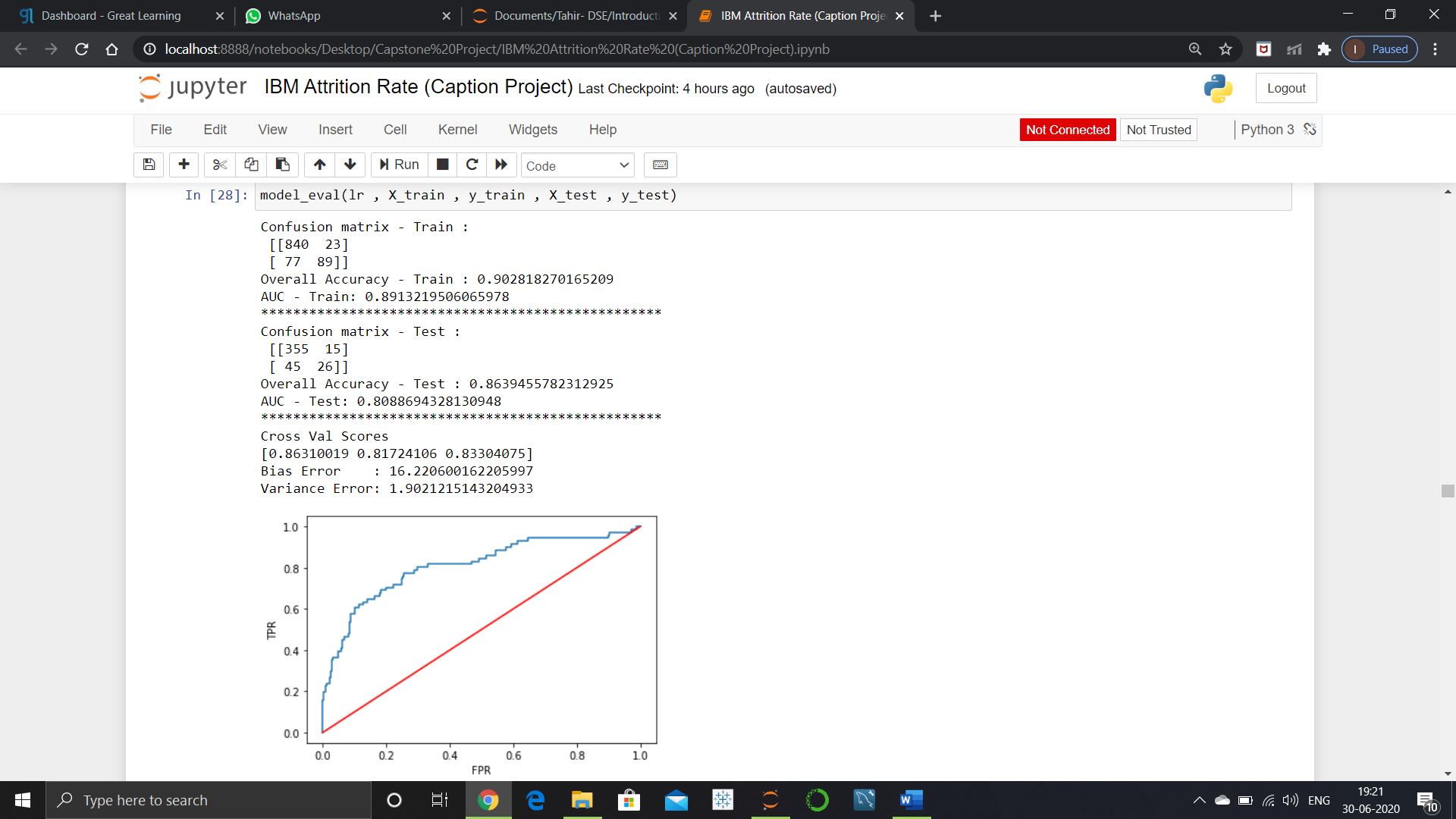
5.ASSUMPTION OF A LARGE SAMPLE SIZE

Finally, logistic regression typically requires a large sample size. A general guideline is that you need at minimum of 10 cases with the least frequent outcome for each independent variable in your model. For example, if you have 5 independent variables and the expected probability of your least frequent outcome is .10, then you would need a minimum sample size of 500 (10\*5 / .10).

Before building any model we have converted the categorical column into dummy columns.

**6.2.1 Logistic Regression :-**

Given below is the performance of Logistic Regression without any feature selection and after which the new scores of logistic regression has improved significantly. From test AUC score 0.80 the model’s test score is optimised to 0.85 after carefully applying recursive feature elimination [RFE]

Logistic Regression without RFE**:**

Logistic Regression with RFE:

Confusion matrix- train:

[[845 17]

[111 56]]

Overall Accuracy- train: 0.8756073858114675

AUC-Train: 0.8302096503049585

Confusion matrix- Test:

[[363 8]

[ 40 30]]

Overall Accuracy- Test: 0.891156462585034

AUC-Test: 0.8523296110897189

Logistic Regression with feature selected using VIF:

Confusion matrix - Train :

[[845 18]

[ 89 77]]

Overall Accuracy - Train : 0.8960155490767736

AUC - Train: 0.8613480573510728

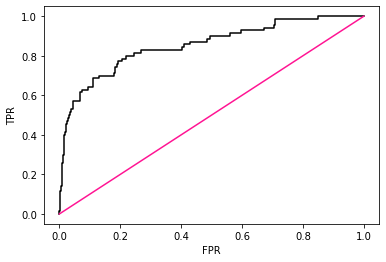
Confusion matrix - Test :

[[353 17]

[ 47 24]]

Overall Accuracy - Test : 0.854875283446712

AUC - Test: 0.7907879710696611



Logistic Regression with feature selected using VIF & SMOTE:

Confusion matrix - Train :

[[791 72]

[ 83 780]]

Overall Accuracy - Train : 0.910196987253766

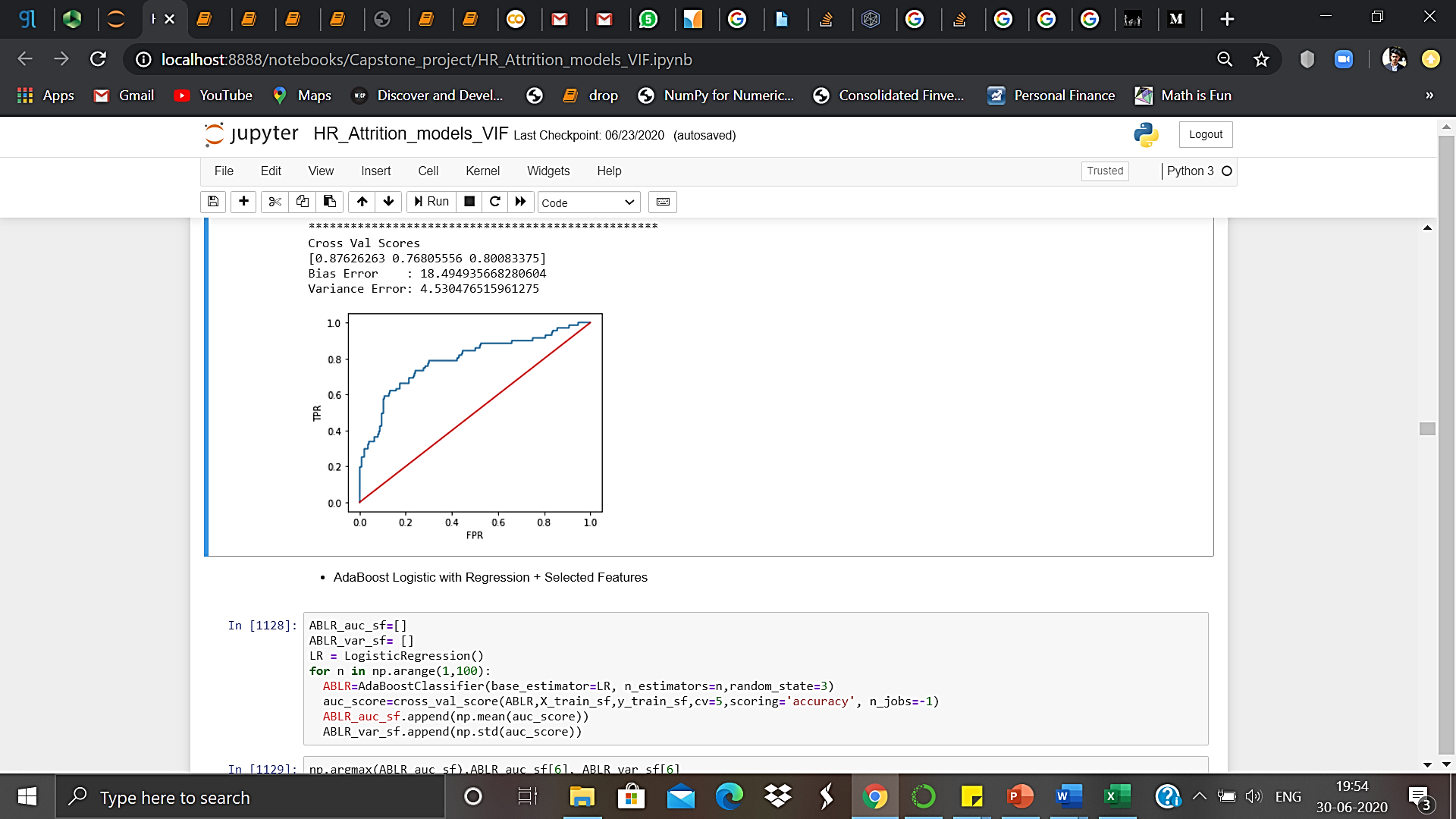
AUC - Train: 0.9614511345128489

Confusion matrix - Test :

[[340 30]

[ 33 38]]

Overall Accuracy - Test : 0.8571428571428571

AUC - Test: 0.7861819566044917

**6.2.2 Naïve Bayes:**

It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

**Gaussian Naïve Bayes:** It is used in classification and it assumes that features follow a normal distribution.

Confusion matrix - Train :

[[591 272]

[ 36 130]]

Overall Accuracy - Train : 0.7006802721088435

AUC - Train: 0.7994178335590334

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Confusion matrix - Test :

[[237 133]

[ 22 49]]

Overall Accuracy - Test : 0.6485260770975056

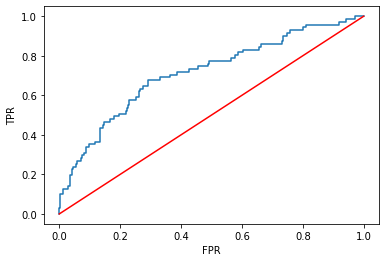
AUC - Test: 0.7134754472782641

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Cross Val Scores

[0.8009794 0.71920909 0.76152638]

Bias Error : 23.942837783732173

Variance Error: 3.3389412731309416

6.3 **Non-Linear Models:**

For non-linear model feature selection is not necessary.

**6.3.1 Decision Tree:**

The below are the some of the assumptions we make while using Decision tree:

At the beginning, the whole training set is considered as the root. Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model. Records are distributed recursively on the basis of attribute values.

Confusion matrix - Train :

[[820 43]

[109 57]]

Overall Accuracy - Train : 0.8522837706511176

AUC - Train: 0.7197783020843513

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Confusion matrix - Test :

[[348 22]

[ 51 20]]

Overall Accuracy - Test : 0.8344671201814059

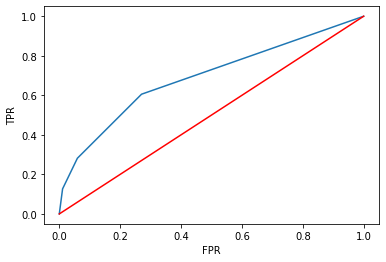
AUC - Test: 0.6899885801294252

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Cross Val Scores

[0.58774523 0.68617758 0.65020481]

Bias Error : 35.86241235229501

Variance Error: 4.06668946773156

**6.3.2. Random Forest:**

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model prediction. After using Randomized Search CV we got best parameter for modelling. Using those hyper parameter Random forest scores are as follows

Confusion matrix- train:

[[862 0]

[105 62]]

Overall Accuracy- train: 0.8979591836734694

AUC-Train: 0.9830987676618921

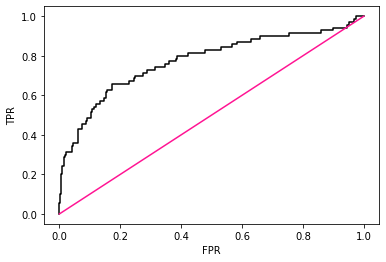
Confusion matrix- Test:

[[369 2]

[ 59 11]]

Overall Accuracy- Test: 0.8616780045351474

AUC-Test: 0.7735078937235272



**6.2.3 KNN:**

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970’s as a non-parametric technique.

Confusion matrix- train:

[[862 0]

[ 0 167]]

Overall Accuracy- train: 1.0

AUC-Train: 1.0

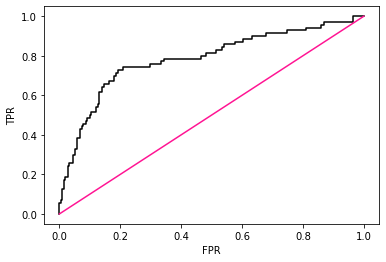
Confusion matrix- Test:

[[371 0]

[ 70 0]]

Overall Accuracy- Test: 0.8412698412698413

AUC-Test: 0.7841740469772814



**6.2.4 SVM:**

The SVM algorithm is implemented in practice using a kernel. The learning of the hyperplane in linear SVM is done by transforming the problem using some linear algebra, which is out of the scope of this introduction to SVM. A powerful insight is that the linear SVM can be rephrased using the inner product of any two given observations, rather than the observations themselves. The inner product between two vectors is the sum of the multiplication of each pair of input values.

Accuracy: 86%

AUC score: 0.83

Variance: 0.0016

**6.3 Ensemble Techniques (Bagging and Boosting):**

**6.3.1 Bagging:**

Bootstrap aggregating, also called bagging (from bootstrap aggregating), is a [machine learning ensemble](https://en.wikipedia.org/wiki/Ensemble_learning) [meta-algorithm](https://en.wikipedia.org/wiki/Meta-algorithm) designed to improve the stability and accuracy of [machine learning](https://en.wikipedia.org/wiki/Machine_learning) algorithms used in [statistical classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression](https://en.wikipedia.org/wiki/Regression_analysis). It also reduces [variance](https://en.wikipedia.org/wiki/Variance) and helps to avoid [overfitting](https://en.wikipedia.org/wiki/Overfitting). Although it is usually applied to [decision tree](https://en.wikipedia.org/wiki/Decision_tree_learning) methods, it can be used with any type of method. Bagging is a special case of the [model averaging](https://en.wikipedia.org/wiki/Ensemble_learning) approach.

**6.3.1.Boosting:**

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), boosting is an [ensemble](https://en.wikipedia.org/wiki/Ensemble_learning) [meta-algorithm](https://en.wikipedia.org/wiki/Meta-algorithm) for primarily reducing [bias](https://en.wikipedia.org/wiki/Supervised_learning#Bias-variance_tradeoff) in [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning), and a family of machine learning algorithms that convert weak learners to strong ones.

**Types of boosting algorithms :**

**6.3.1.1 AdaBoost (Adaptive Boosting):**

It fits a sequence of weak learners on different weighted training data. It starts by predicting original data set and gives equal weight to each observation. If prediction is incorrect using the first learner, then it gives higher weight to observation which have been predicted incorrectly. Being an iterative process, it continues to add learner(s) until a limit is reached in the number of models or accuracy. Mostly, we use decision stamps with AdaBoost. But, we can use any machine learning algorithms as base learner if it accepts weight on training data set. We can use AdaBoost algorithms for both classification and regression problem. Default base estimator is decision tree classifier.

**6.3.1.2 Gradient Tree Boosting (Gradient Boosting):**

In gradient boosting, it trains many model sequentially. Each new model gradually minimizes the loss function (y = ax + b + e, e needs special attention as it is an error term) of the whole system using [Gradient Descent](https://en.wikipedia.org/wiki/Gradient_descent) method. The learning procedure consecutively fit new models to provide a more accurate estimate of the response variable. The principle idea behind this algorithm is to construct new base learners which can be maximally correlated with negative gradient of the loss function, associated with the whole ensemble.

Accuracy: 85%

AUC score: 0.80 (0.0010)

**6.3.1.3 XGBoost:**

Xgboost stands for extreme Gradient Boosting and is developed on the framework of gradient boosting. It boosts the performance of a regular gradient boosting model. XGBoost used as a more regularized model formalization to control over-fitting, which gives it better performance.

The two reasons to use XGBoost are also the two goals of the project:

1. Execution Speed.
2. Model Performance.

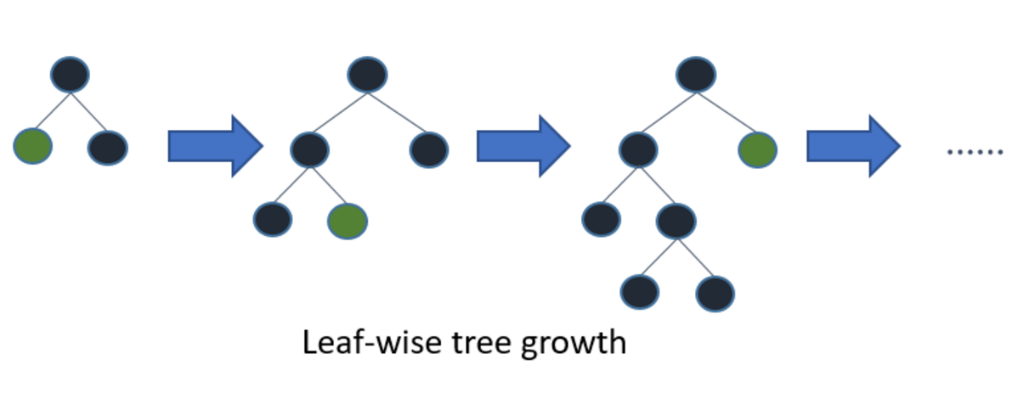
Accuracy: 87.07%

AUC score: 0.804 (0.041)

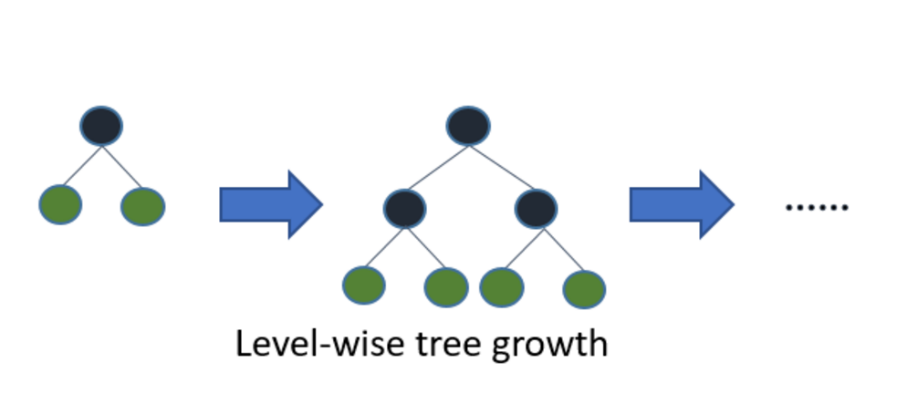
**LightGBM(LGBM):**

LightGBM is a gradient boosting framework that uses tree based learning algorithms. **LightGBM grows tree vertically**while other algorithm grows trees horizontally meaning that Light GBM grows tree **leaf-wise**while other algorithm grows level-wise. It will choose the leaf with max delta loss to grow. When growing the same leaf, Leaf-wise algorithm can reduce more loss than a level-wise algorithm.

Below diagrams explain the implementation of LightGBM and other boosting algorithms.



How LGBM works



How other boosting algorithm works

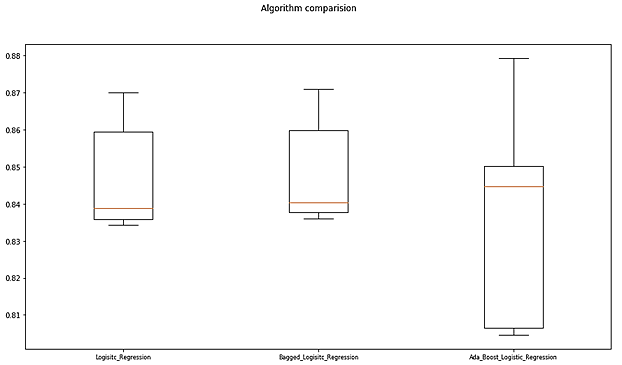
LGBM is designed to be distributed and efficient with the following advantages:

1. Faster training speed and higher efficiency
2. Lower memory usage.

Accuracy: 86%

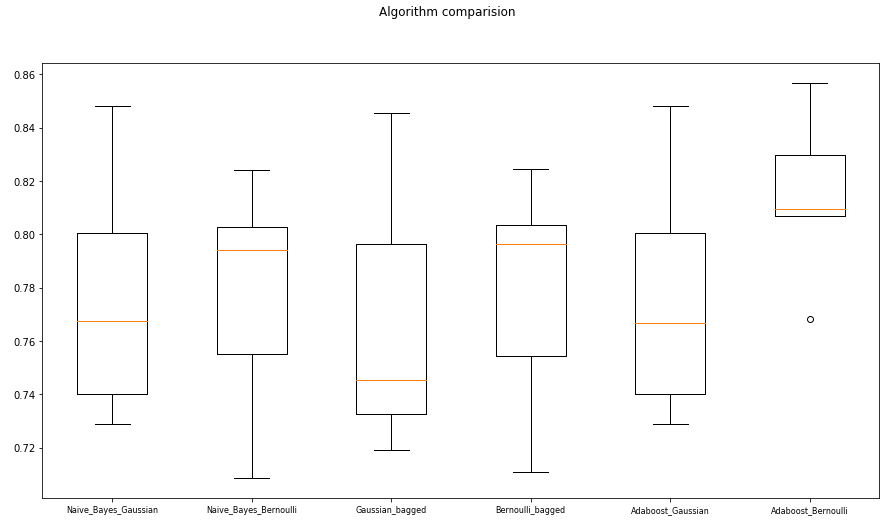
AUC score: 0.76(0.046)

**6.4 Comparison of all liner model**



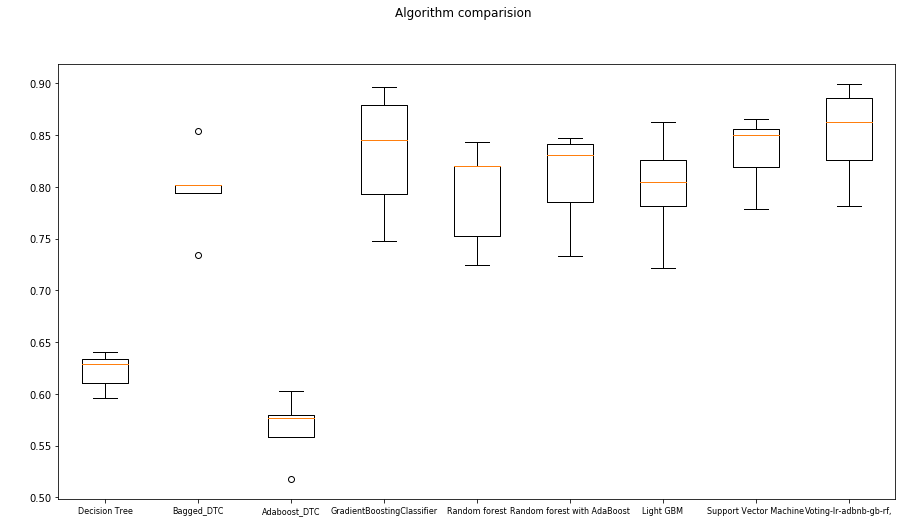
|  |  |  |
| --- | --- | --- |
| **Algorithms** | **Bias Error(%)** | **Variance Error(%)** |
| Logistic Regression | 15.231 | 1.608 |
| Bagged\_Logisitc\_Regression | 15.102 | 1.555 |
| AdaBoost\_Logistic\_Regression | 16.296 | 3.167 |

|  |  |  |
| --- | --- | --- |
| **Algorithms** | **Bias Error(%)** | **Variance Error(%)** |
| Naive\_Bayes\_Gaussian | 22.302 | 4.838 |
| Naive\_Bayes\_Bernoulli | 22.313 | 4.564 |
| Gaussian\_bagged | 23.215 | 5.237 |
| Bernoulli\_bagged | 22.218 | 4.528 |
| Adaboost\_Gaussian | 22.317 | 4.842 |
| Adaboost\_Bernoulli | 18.578 | 18.578 |

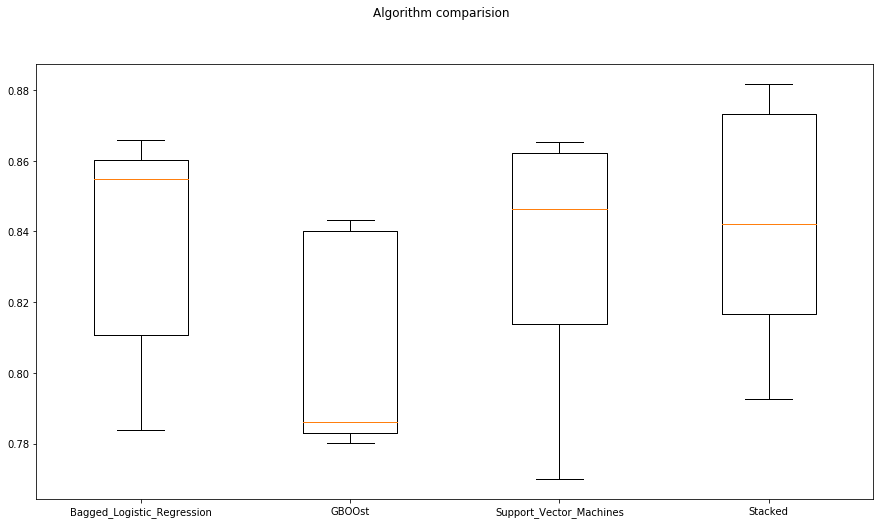


**Comparison of Non-Linear Models:**

|  |  |  |
| --- | --- | --- |
| **Algorithms** | **Bias Error(%)** | **Variance Error(%)** |
| Decision Tree | 38.435 | 2.585 |
| Bagged\_Decision\_Tree | 21.307 | 4.06 |
| AdaBoost\_Decision\_Tree | 43.311 | 3.166 |
| GradientBoostingClassifier | 16.75 | 6.1546 |
| Random forest | 20.771 | 5.084 |
| Random forest with AdaBoost | 19.239 | 4.817 |
| Light GBM | 20.056 | 5.274 |
| Support Vector Machine | 21.6241 | 2.464 |
| Voting(LR, Ad\_BNB, GB, Ad\_rf) | 14.898 | 4.664 |



**Comparison of the best models:**



**7. Limitations**

Since our data set had very few records, we were confined to create a model within that. In future, the dataset can include more detailed data so that we can dive deeper and based on the attrition rate, we can give benefits to the employees so that the employee attrition rate gets decreased.

8. **Conclusion**:

The study has set out to be a real time application in the organizations where the management can predict the future actions of the employees based on their

records and observations. The main focus was to build a model which can

efficiently predict the employees that might leave the company in future, and

considering the real scenario, the higher management will be more interested to know the potential employees who might actually leave so that they can set

their attention on them to stop them to do so. With the use of RFE and VIF base feature selection technique, this research is able to figure out the major

reasons for the turnover.

After applying various models, it was observed that Bagged Logistic Regression with RFE gave the best AUC score of 84 % and variance 0.0012 in the testing phase though other models performed well in the training. Hence, it can be

concluded that the Logistic Regression with RFE feature selection are more

reliable than other methods of feature selection as well as other algorithms to

increase employee retention.

Hence, with the help of our model, companies can now have a smooth flow of

human talent without any major hiccups. All factors are immensely scrutinized and not only will companies benefit but employees too. Companies have the

ability to forecast and take measures well in advance while smarter retention

policies , ESOPs and incentives can be formulated for their respective

employees. **It’s a win-win situation all round.**

