

# **EMPLOYEES RETENTION ANALYSIS**

**TEAM NAME:** Team StackOverflow

**TEAM NUMBER:** 10

**TEAM MEMBERS:** Prashant Channappa Mural,  
Shashank Telkhade,  
Shikhar Agrawal,  
Sanket Waghmare

## **Introduction**

Employee attrition is the gradual reduction in employee numbers over the course of one-year. Attrition is a problem that impacts all businesses, industries, and the size of the company. Employee attrition leads to increased costs for a business, including the cost of business disruption, hiring and training inexperienced staff. As such, there is great business interest in understanding the drivers of, and minimizing staff attrition.

The objective of our project is

- To analyse numerous factors responsible for the ATTRITION of employees
- Finding efficient ways to retain the productive and talented employees

## **Methods**

The implementation methodology constitutes the following:

1. Data Pre-processing
2. Data Visualisation
3. Building Decision Tree Classifier
4. Analysis of Factors Responsible for Attrition
5. Analysis of Factors to Retain an Employee

## **Data Pre-Processing**

### **Dataset**

The data set represents employee reviews from Glassdoor with target value 'Attrition'.

- 'Yes' - Resign (+ve)
- 'No' - Doesn't Resign (-ve)

Features include Numerical and Categorical values.

Some level of attrition will always be there, but the company wants to minimize this and keep the best employee from leaving.

The data set is used to run analysis of employees, to find certain “at risk” categories of employees.

The dataset consists of 21 columns out of which we have selected 14 columns based on our intuition. This is done because the columns like Sr No., emp id will not have any significant effect on our analysis

Also, we added 3 columns for our analysis

- Polarity – This shows the sentiment of the ‘text reviews’ which we have in the dataset. The data is converted to a numerical format for analysis using the TextBlob library. The polarity value ranges from -1 to 1
- Subjectivity - This shows the subjectivity of each review how much significance it holds.
- NormalizedMarketValuation - For the column CompanyName, the unique companies in the dataset added a new column Marketvaluation. Normalized the values based on maximum

10) → Not considered (No views)

Data: 650 → 549 (taken) → (-35) → (-12) → (-9)

```
outlier
"C:\Users\dsp lab\anaconda3\envs\e9_241_dip\python.exe" "C:/Users/dsp lab/PycharmProjects/pythonProject2/mini_project/outlier.py"
shape of the dataframe is (650, 25)
VALID NLP : number of decimal values in subjectivity 549
NEUTRAL : number of rows where polarity is between (-0.1, 0.1) 35
NON NEUTRAL : number of rows in df_non_neutral 514
FALSE NEGATIVE : rows where polarity is between (-0.5, -1) and Attrition is No 12
FALSE POSITIVE : rows where polarity is between (0.75, 1) and Attrition is Yes 9
shape of df_filtered (594, 25)
Process finished with exit code 0
```

650  
549  
-35 → Neutral  
-12  
-9  
-56 (filtered)  
594 (Remaining)

→ Basis for Data Filtering

→ Reviews: Abstract text but capture general sentiment of an employee towards organization.

→ NLP: Applied on Reviews to convert into Subjectivity Polarity

→ Subjectivity: Gives the extent to which the review was sentimental.

→ Polarity: Gives the sentimental value of employee

→ Filtering logic

→ Consider Only 'Reviews' to filter data: as it adequately represents overall sentiment of employee

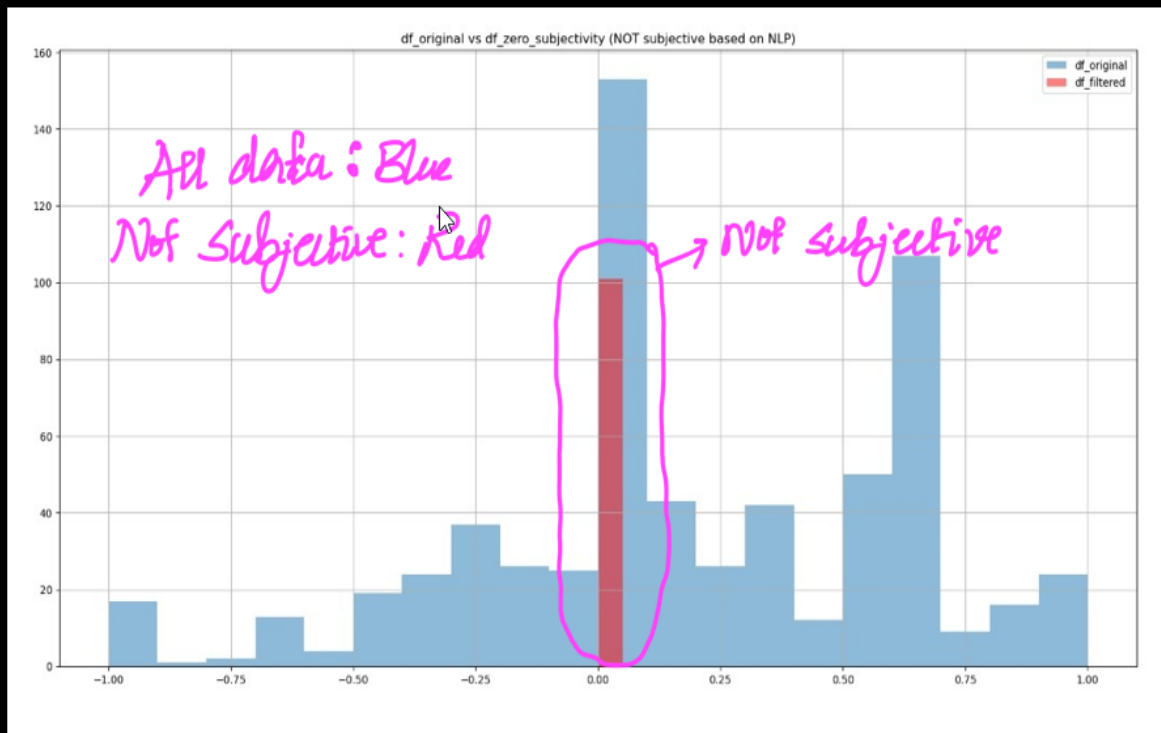
→ Take Only non-zero Subjective data: Since if it's zero then review does not have adequate info.

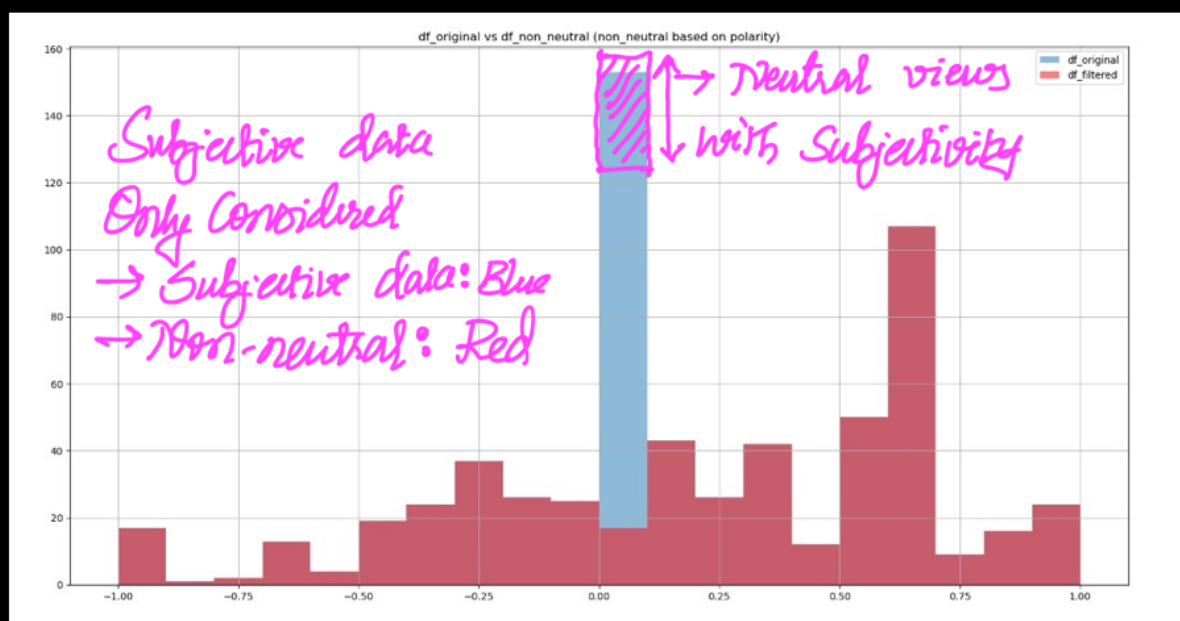
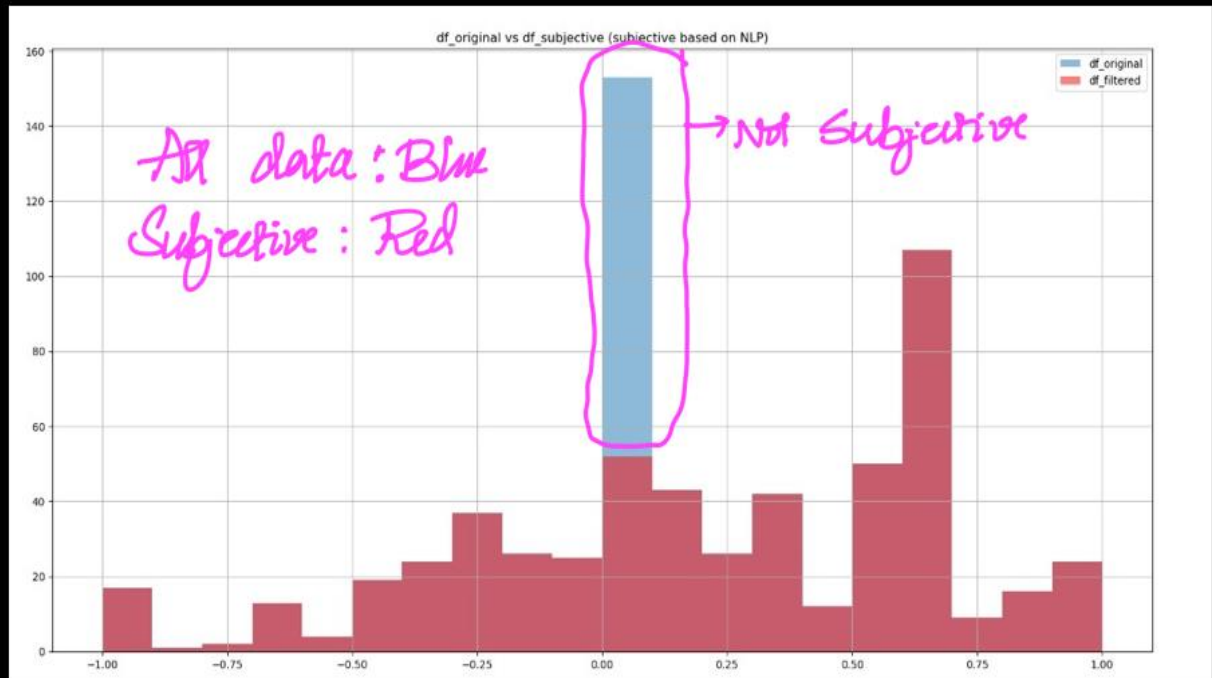
→ Drop Neutral Data: Data around centre which is neither -ve nor +ve

→ Drop False +ves & -ves: Contradictory data points.

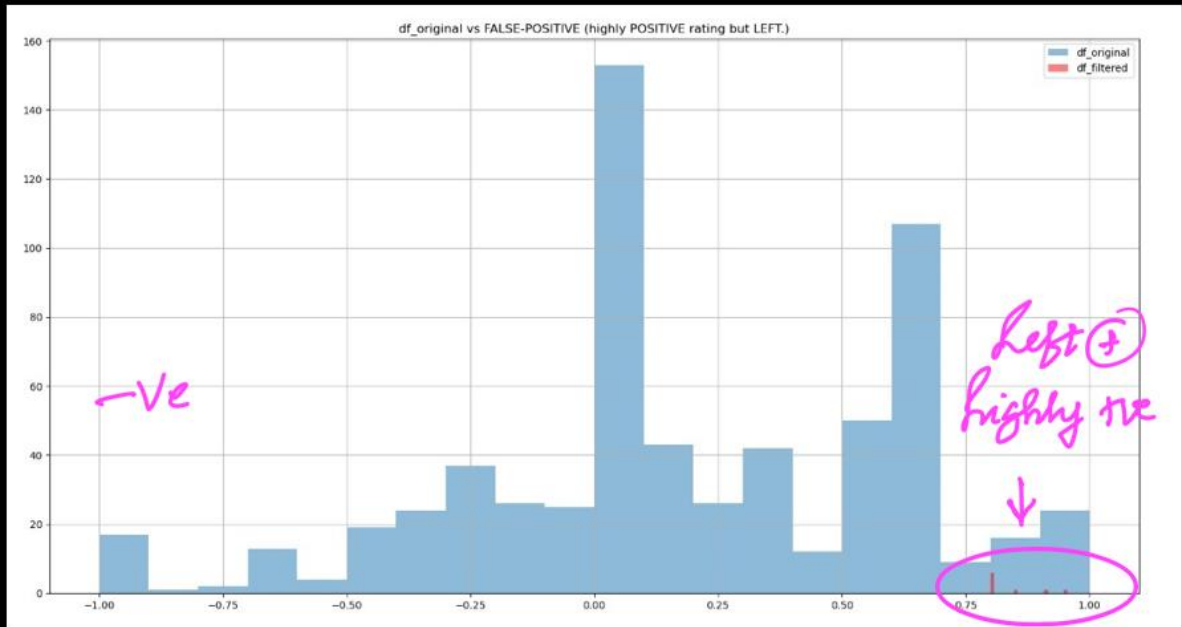
→ Bias Removal ← Data Filtering

→ Fraction of data : Not Subjective (Red)





→ False Positives

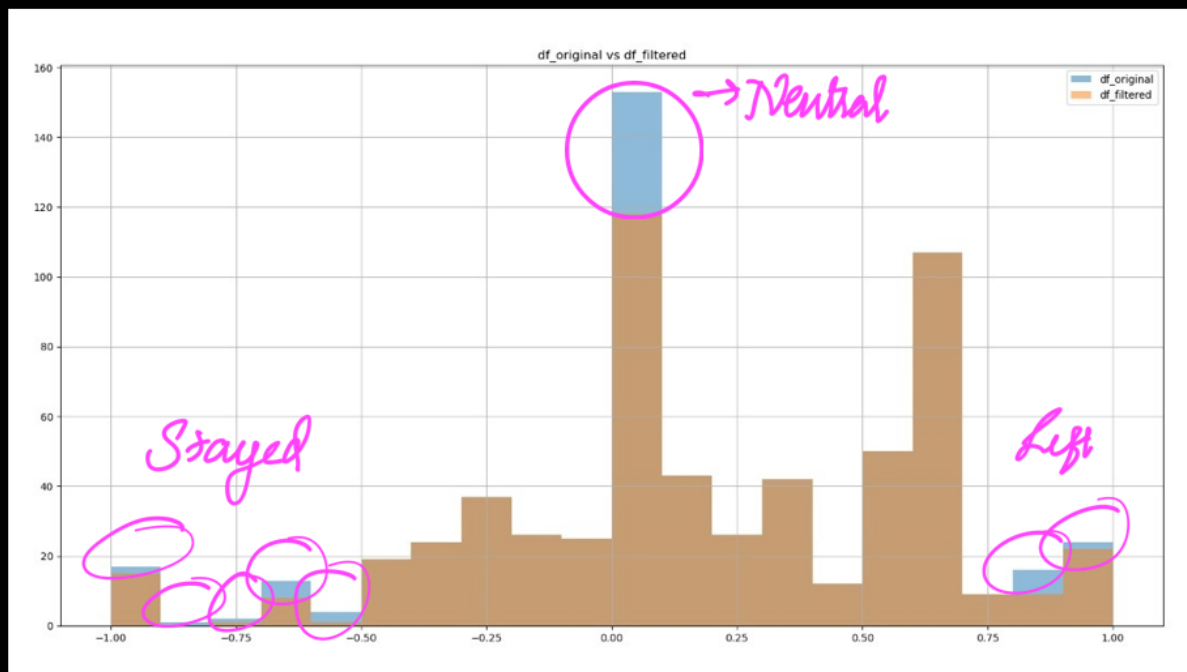


★ highly +ve views (but still) left company

→ False Negatives

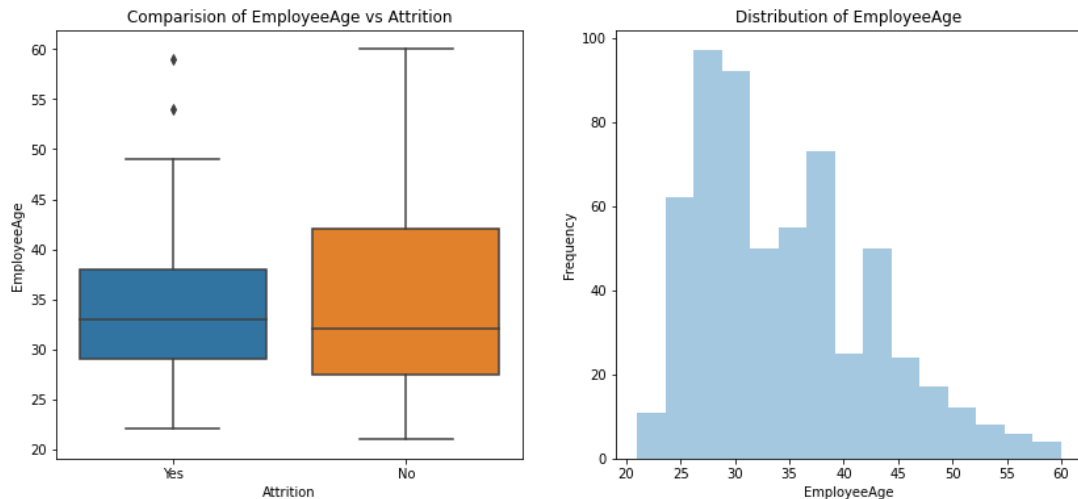


Highly -ve Sentiments ( $\frac{\text{but}}{\text{still}}$ ) chose to stay back



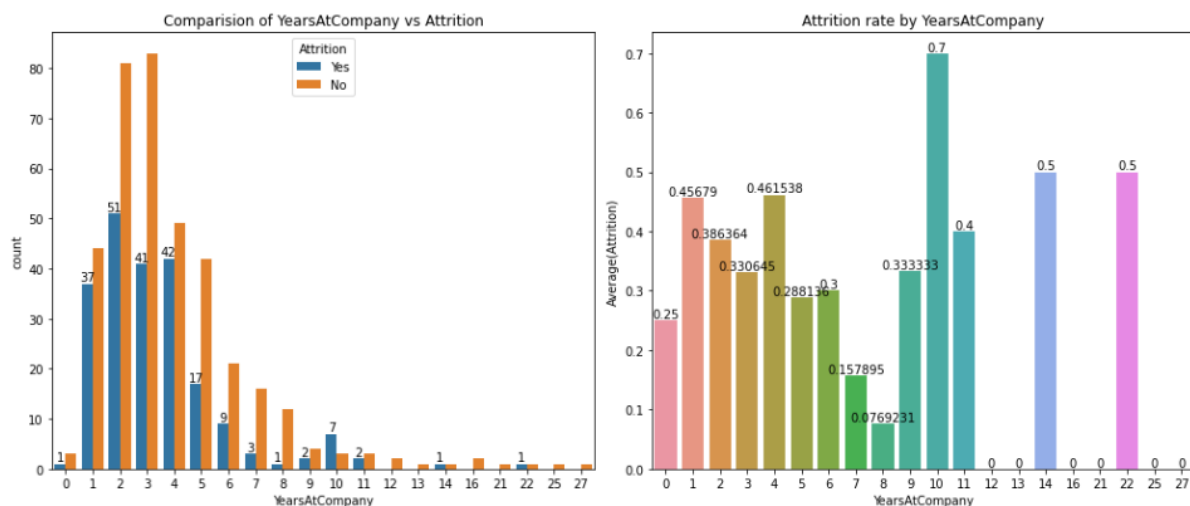
The data obtained in step 1 was used to carry out data visualization. Various Graphs were plotted between the features and the Target Variable i.e., 'Attrition' to visually see the correlation between the features and 'Attrition' and also to draw various possible inferences from these plots. Also, some features contain ordinal values with 1 as the lowest value and 5 as the highest value. Shown below are the various plots obtained:

(i) **Employee Age vs Attrition**



- It was observed that the median of the age of employee's is 30-40 years, with minimum and maximum being 21 and 60 years, respectively.
- From the above box plot, it is observed that the majority of employees that left the company falls below the age of 38 years, and the employees that didn't leave the company are of age 28 to 42 years.

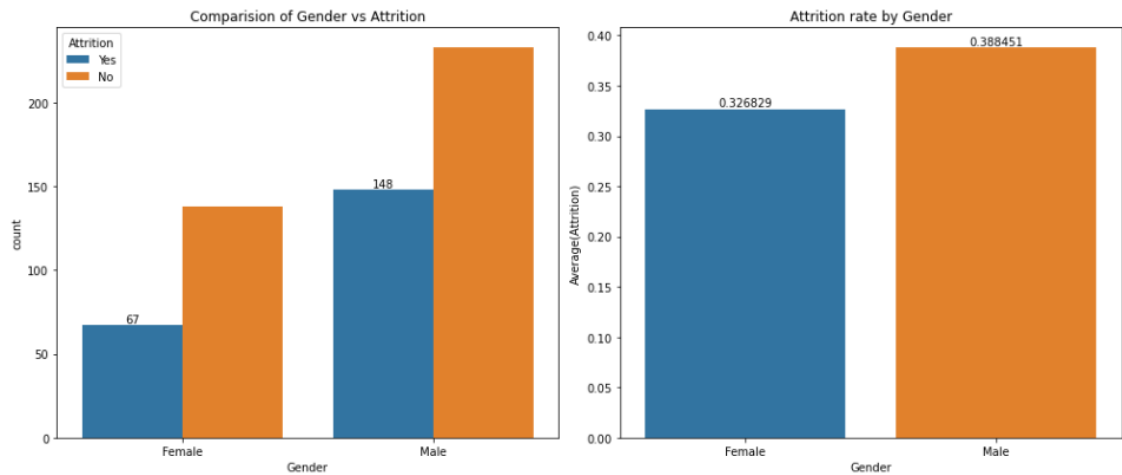
(ii) **Years At Company vs Attrition**



- From the above plot, it can be observed that the highest percentage of attrition is amongst people who've spent 10 years (7 employees out of 10) in the company which constitutes only a small number.
- The attrition is maximum amongst employees who've spent years in the company between 1 to 7 years.

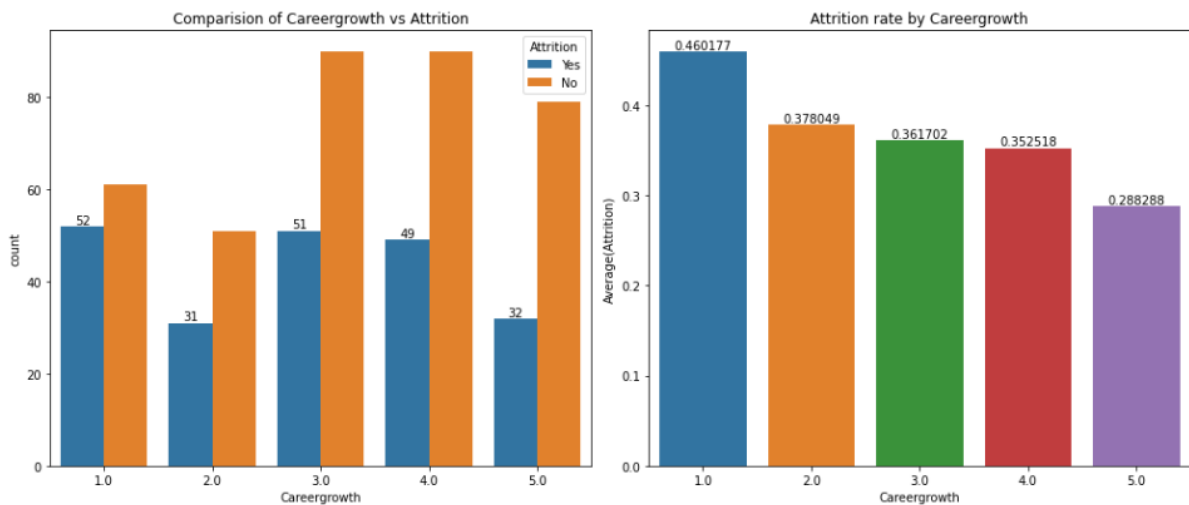


(iii) **Gender Vs Attrition**



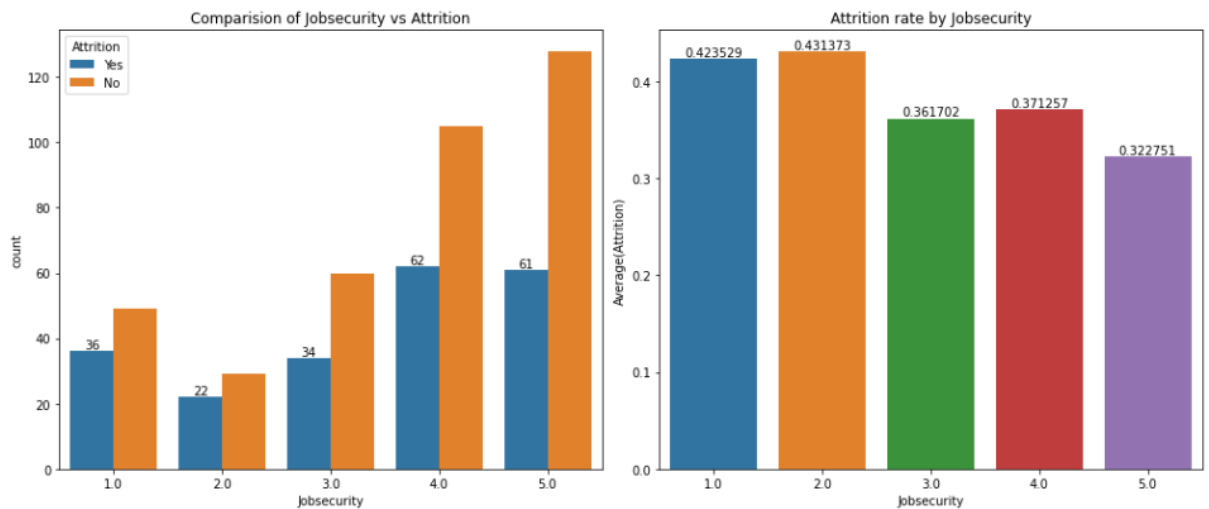
- From the above plots, it can be observed that the percentage of attrition amongst 'Male' gender is roughly 39 % compared to the females with 33 % approximately.

(iv) **Career Growth Vs Attrition**



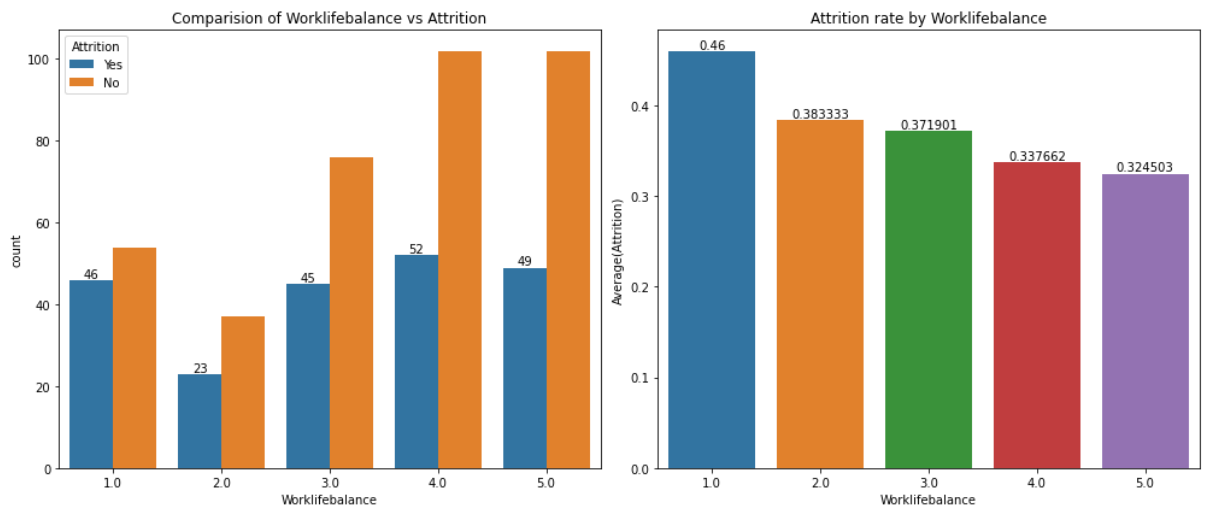
- Career Growth is an ordinal value, where 1 represents an exceptionally low and 5 represents an extremely high career growth.
- From the plots we can see that the highest amount of attrition (approx. 46%) is amongst employees who have found career growth as lowest, and it the attrition reduces with the improvement in the feedback on career growth.

(v) **Job Security Vs Attrition**



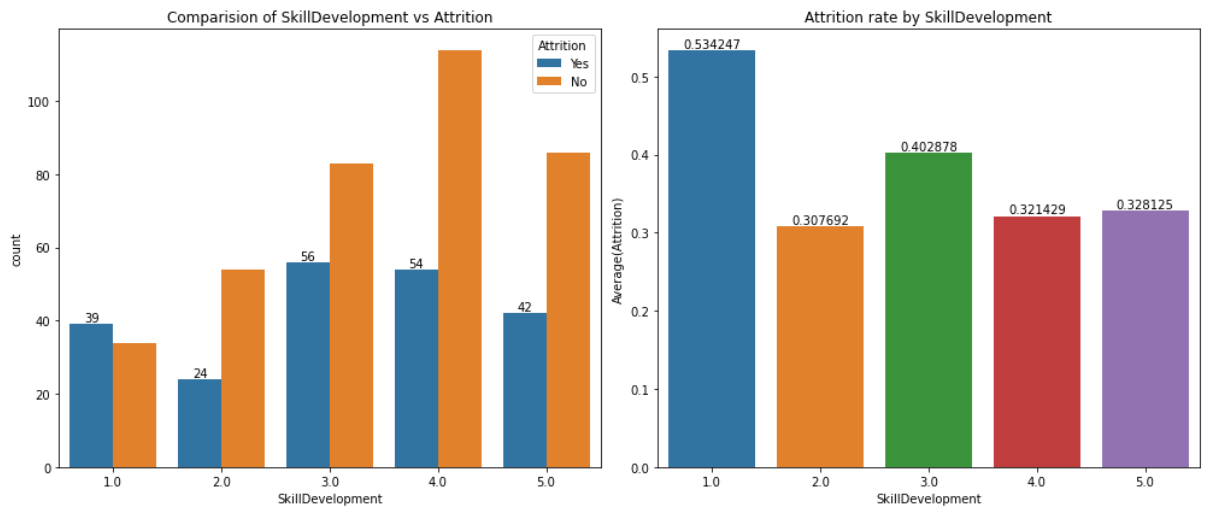
- Job Security is an ordinal value, where 1 represents an exceptionally low and 5 represents an extremely high Job Security.
- From the above plots, it is observed that Highest amount of 'Attrition' of 43% and 42%, is amongst the employees who gave feedback on job security as 0 and 1 respectively which improves as the feedback on Job Security improves.

(vi) **Work Life Balance Vs Attrition**



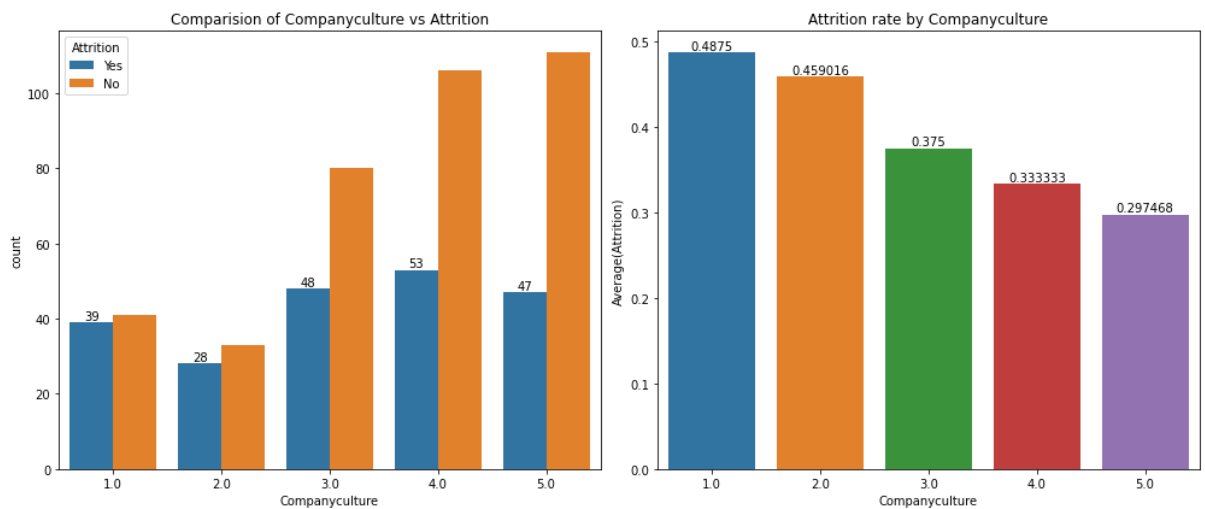
- Work Life Balance is an ordinal value, where 1 represents an exceptionally low and 5 represents an extremely high Work Life Balance.
- 46 % employees left their job who rated their work life balance as poor. This percentage drops with an improvement in the feedback received on work life balance.
- 32 % attrition is noted amongst employees who found the best work life balance.

(vii) **Skill Development Vs Attrition**



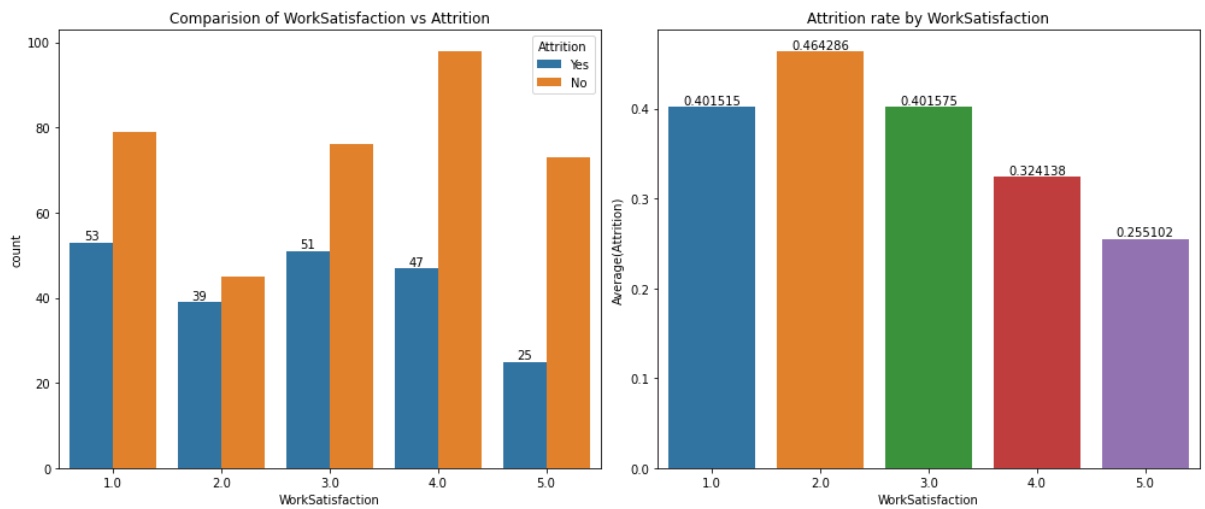
- Skill Development is an ordinal value, where 1 represents an exceptionally low and 5 represents an extremely high Skill Development
- From the above plots, more than 53 % attrition is observed in employees who found extremely poor skill development.

(viii) **Company Culture Vs Attrition**



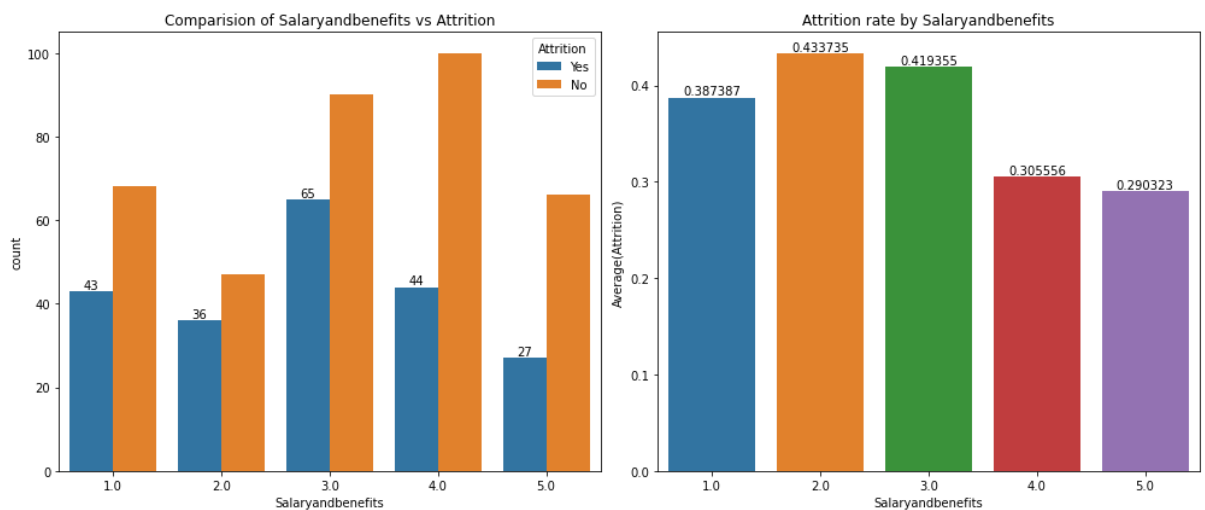
- Company Culture is an ordinal value, where 1 represents an extremely poor and 5 represents a particularly good Company Culture
- Attrition is highest, roughly 48% amongst employees who found extremely poor company culture, which drops as the company culture improves.

(ix) **Work Satisfaction Vs Attrition**



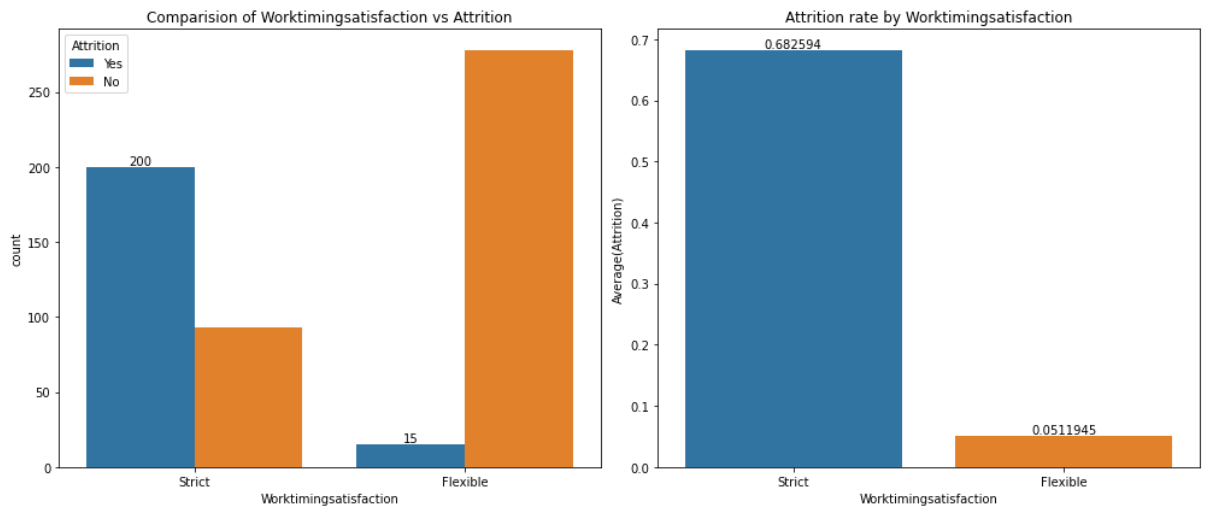
- Work Satisfaction is an ordinal value, where 1 represents an extremely poor and 5 represents a particularly good Work Satisfaction.
- Higher percentages of attrition observed amongst employees which rated work satisfaction between average to poor.

(x) **Salary & Benefits Vs Attrition**



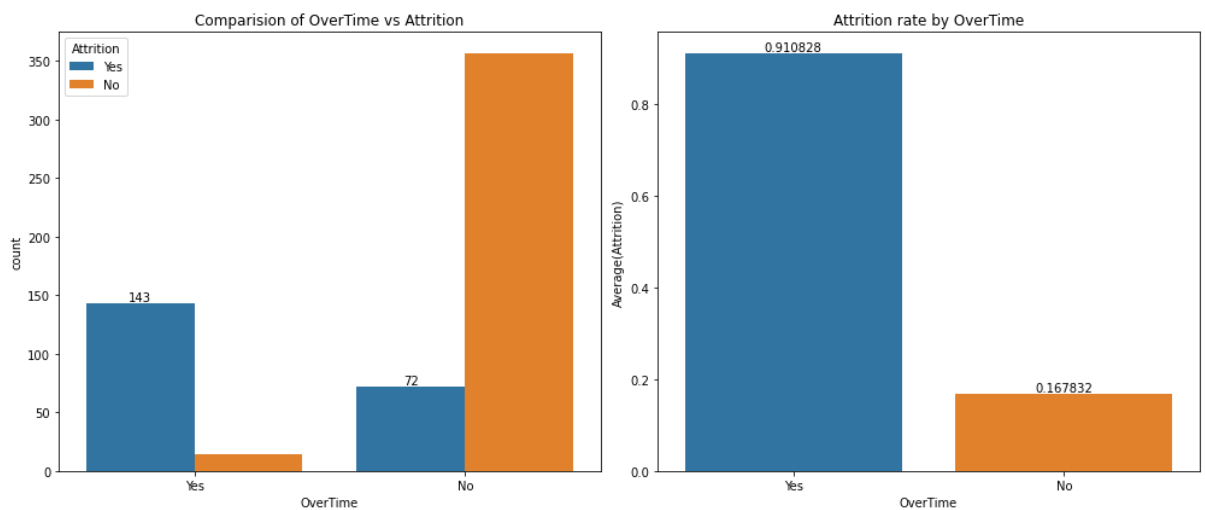
- Salary and Benefits is another ordinal value where 1 represents an extremely poor and 5 represents particularly good Salary & Benefits.
- Attrition has dropped to 30% and 29% where the salary & benefit is good and best respectively as compared to the attrition which was observed to be highest amongst employees who found Salary & Benefits between extremely poor to average.

(xi) **Work Timing Satisfaction Vs Attrition**



- Approximately 69 % of employees who have experienced 'Strict' Work Timings have left the company and Attrition amongst employees with 'Flexible' work time is observed as 5 % only.
- Therefore, it is also a strong indicator of Attrition.

(xii) **Overtime Vs Attrition**

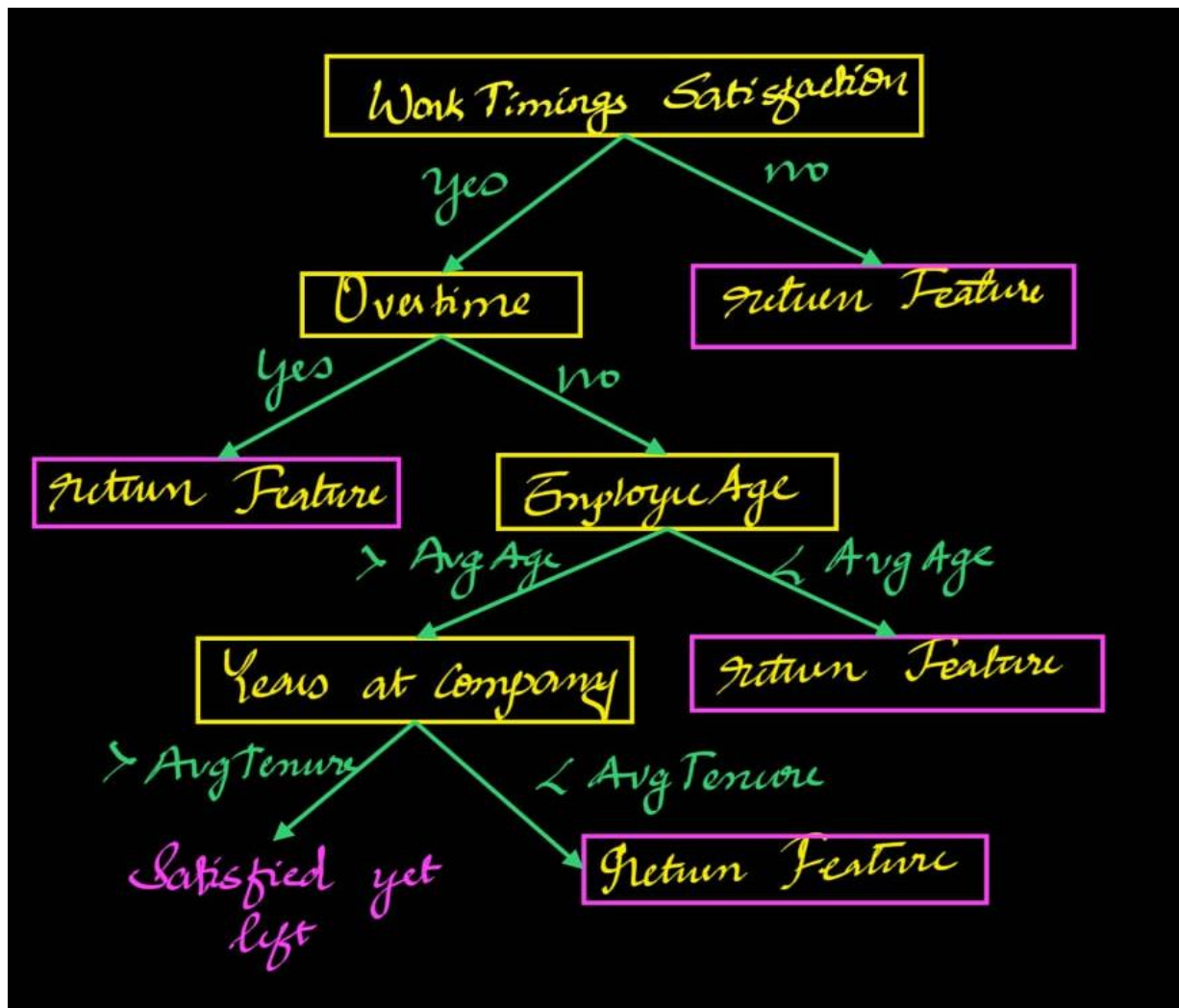


- More than 91% of employee's who worked overtime has left the company, whereas approx 17% of employee's who have experienced overtime has not left the company. Therefore, overtime is a strong indicator of attrition.

### Building Decision Tree Classifier

As the Attrition is the decision taken by the employee based on the challenges faced by him at his/ her firm. We have used Decision Tree Classifier for predicting with what probability employee may leave the firm, also termed as the Risk predicted.

### Flow of the Decision Tree



### RESULTS

#### Metrics for evaluation:

#### Confusion Matrix

Actual vs Predicted	No	Yes
No	61	6
Yes	3	35

**F1 Score:** 0.9247

**Analysis of Factors Responsible for Attrition**


Random Samples of Employees.	Risk of Attrition	Reasons	Ground Truth
1	0.93	Work Time Satisfaction	1
2	0.33	None	0
3	0.72	Over Time	1
4	0.93	Work Time Satisfaction	1
5	0.93	Work Time Satisfaction	1


```
er.py x data_w_sentiments.csv x
outlier
"C:\Users\dsp lab\anaconda3\envs\e9_241_dip\python.exe" "C:/Users/dsp lab/PycharmProjects/pythonProject2/mini_project/outlier.py"
shape of the dataframe is (650, 25)
VALID NLP : number of decimal values in subjectivity 549
NEUTRAL : number of rows where polarity is between (-0.1, 0.1) 35
NON NEUTRAL : number of rows in df_non_neutral 514
FALSE NEGATIVE : rows where polarity is between (-0.5, -1) and Attrition is No 12
FALSE POSITIVE : rows where polarity is between (0.75, 1) and Attrition is Yes 9
shape of df_filtered (594, 25)

Process finished with exit code 0
|
```

★ Method to Get Optimal Centre & Threshold

→ Data was filtered with Centres  $[-0.2, -0.1, \dots, 0.5]$   
with various thresholds  $[0.1, 0.2, 0.3]$

$\Rightarrow$  Neutral values = 

$\Rightarrow$  False  $\left| \begin{array}{l} -ve \\ +ve \end{array} \right.$  =   
highly -ve reviews but still stayed      highly +ve reviews but still left

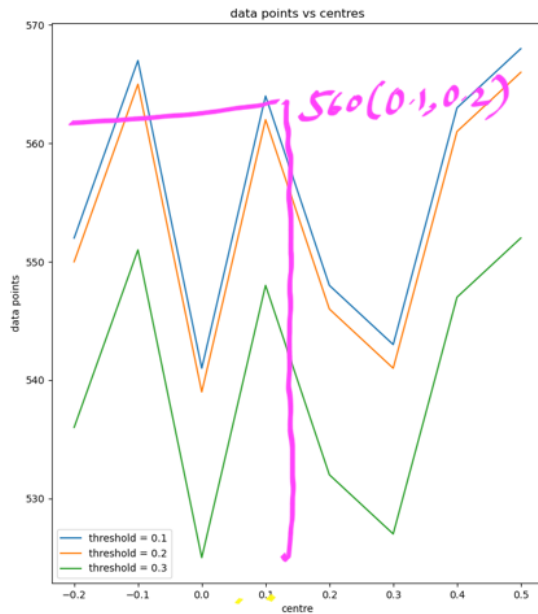
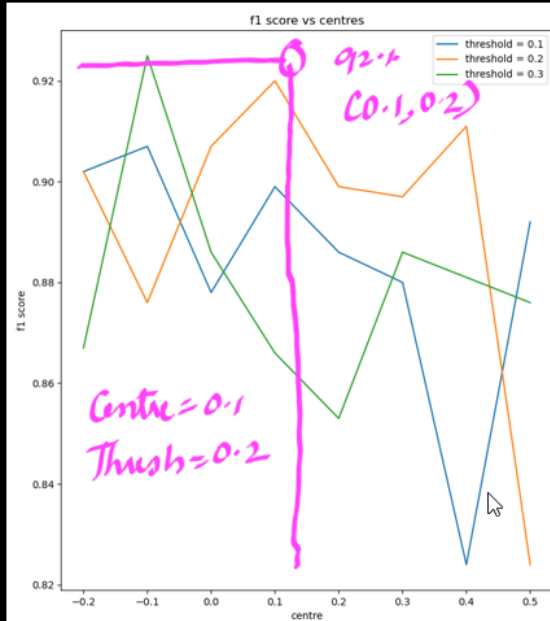
NOTE: The final requirement of this analysis would be to choose those values of Centre & threshold that gives us the higher accuracy as well as considers adequate datapoints.

★ Plots below are Obtained for Accuracy Vs Datapoints

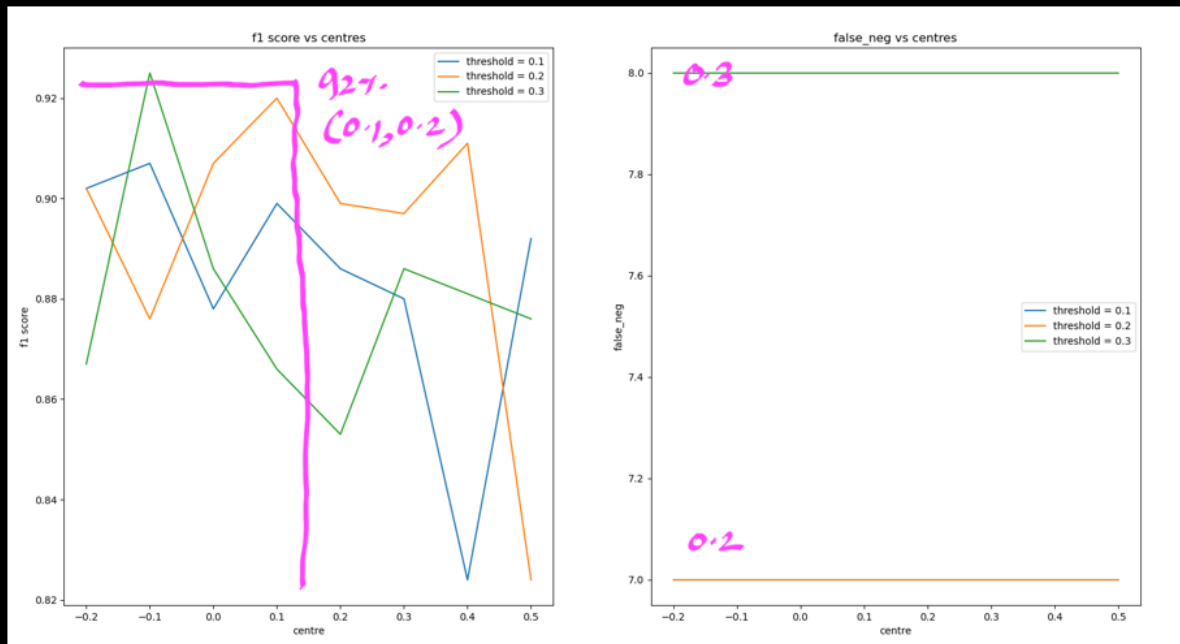
★ When all the features are considered.



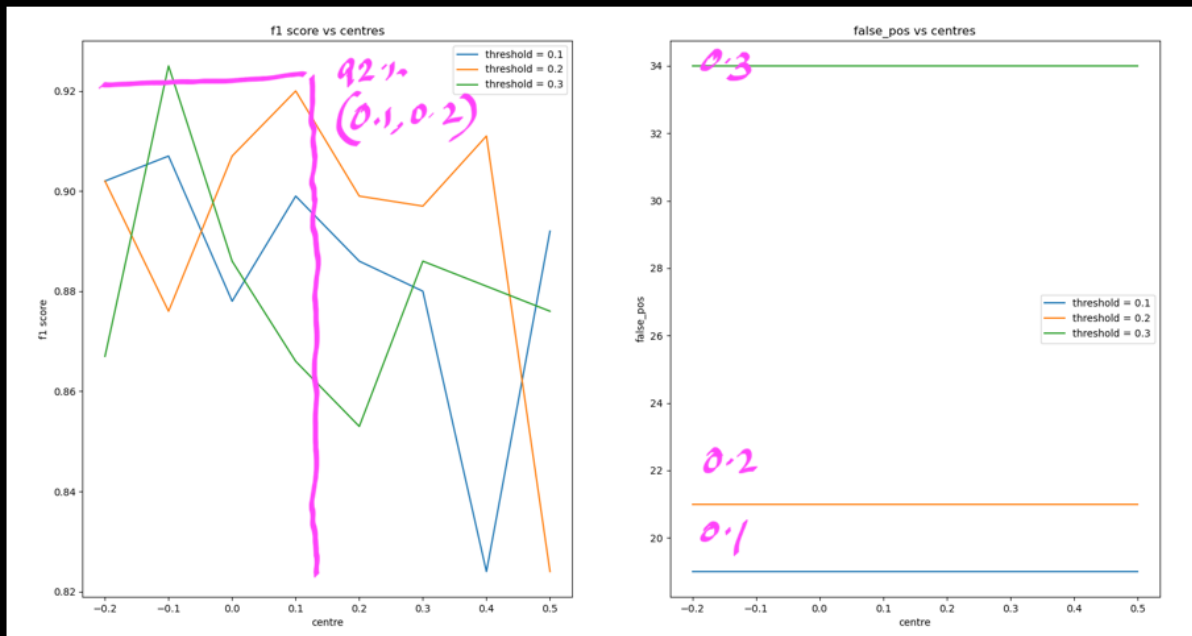
# → Accuracy (vs) Data-points considered



→ Accuracy V/s False-negative dropped



→ Accuracy V/s False-positives dropped

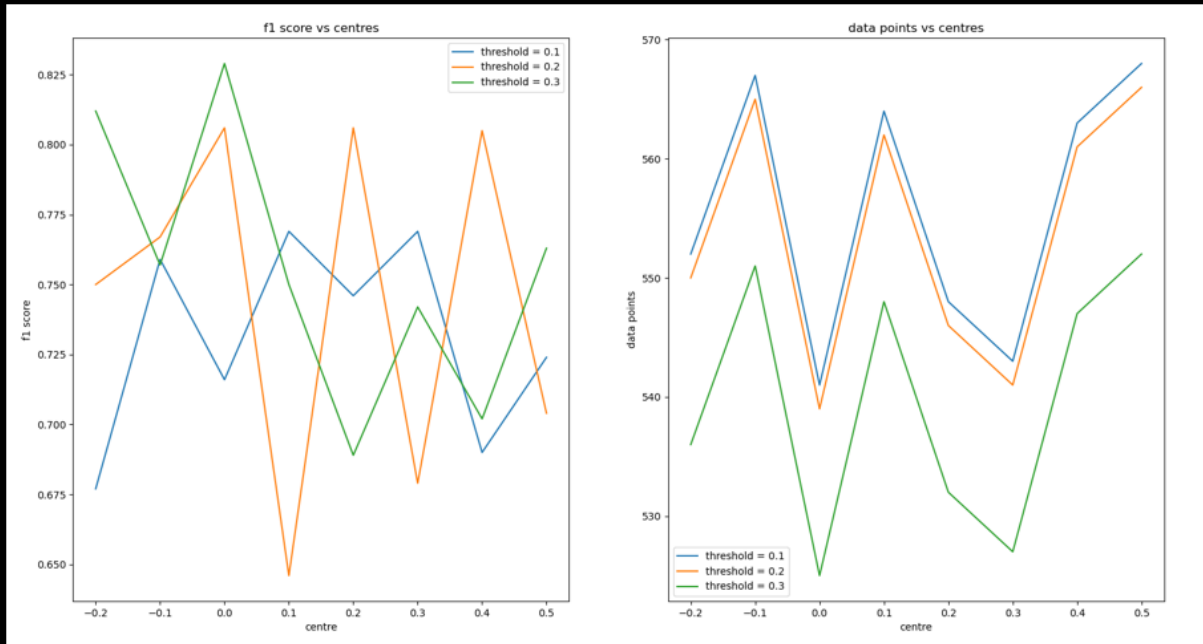


## → Accuracy and related data.

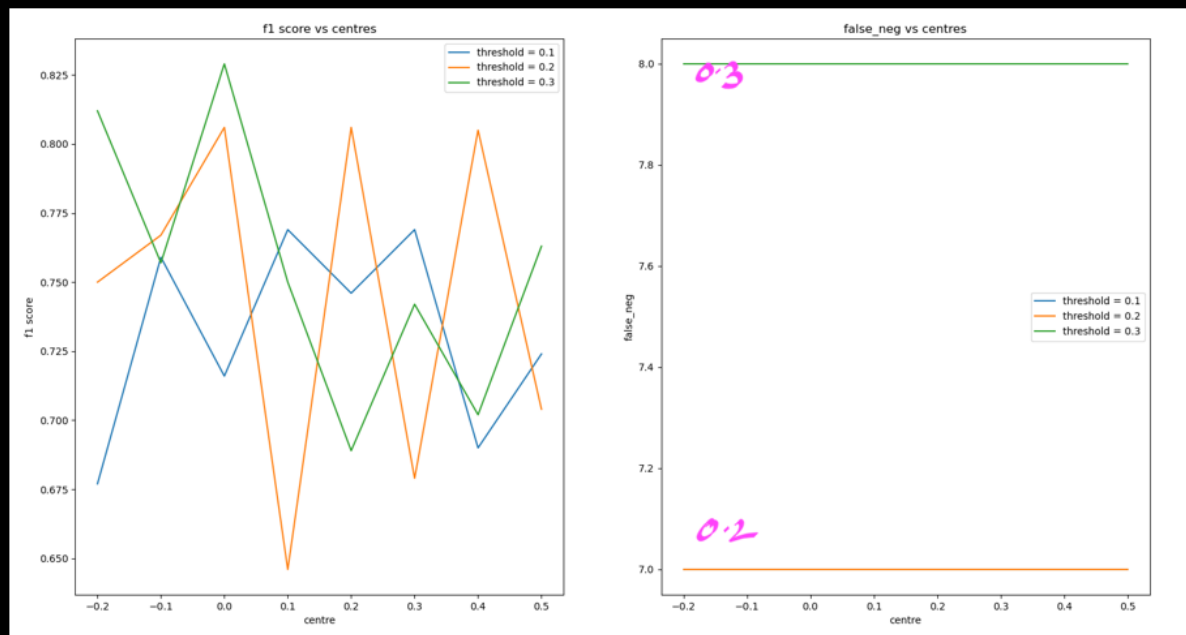
```
data points [552, 567, 541, 564, 548, 543, 563, 568, 550, 565, 539, 562, 546, 541, 561, 566, 536, 551, 525, 548, 532, 527, 547, 552]
f1 scores [0.902, 0.907, 0.878, 0.899, 0.886, 0.88, 0.824, 0.892, 0.902, 0.876, 0.907, 0.92, 0.899, 0.897, 0.911, 0.824, 0.867, 0.925, 0.886, 0.866, 0.853, 0.886, 0.881, 0.876]
neutral [66, 51, 75, 53, 69, 74, 54, 48, 66, 51, 75, 53, 69, 74, 54, 48, 66, 51, 75, 53, 69, 74, 54, 48]
false negative [7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7]
false positive [19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19]
length of f1_scores 24
length of centres 24
length of thresholds 24
length of false_pos 24
length of false_neg 24
length of neutral 24
length of data_points 24
```

	f1_scores	centres	thresholds	false_pos	false_neg	neutral	data_points
0	0.902	-0.2	0.1	19	7	66	552
1	0.907	-0.1	0.1	19	7	51	567
2	0.878	0.0	0.1	19	7	75	541
3	0.899	0.1	0.1	19	7	53	564
4	0.886	0.2	0.1	19	7	69	548
5	0.880	0.3	0.1	19	7	74	543
6	0.824	0.4	0.1	19	7	54	563
7	0.892	0.5	0.1	19	7	48	568
8	0.902	-0.2	0.2	21	7	66	550
9	0.876	-0.1	0.2	21	7	51	565
10	0.907	0.0	0.2	21	7	75	539
11	0.920	0.1	0.2	21	7	53	562
12	0.899	0.2	0.2	21	7	69	546
13	0.897	0.3	0.2	21	7	74	541
14	0.911	0.4	0.2	21	7	54	561
15	0.824	0.5	0.2	21	7	48	566
16	0.867	-0.2	0.3	34	8	66	536
17	0.925	-0.1	0.3	34	8	51	551
18	0.886	0.0	0.3	34	8	75	525
19	0.866	0.1	0.3	34	8	53	548
20	0.853	0.2	0.3	34	8	69	532
21	0.886	0.3	0.3	34	8	74	527
22	0.881	0.4	0.3	34	8	54	547
23	0.876	0.5	0.3	34	8	48	552

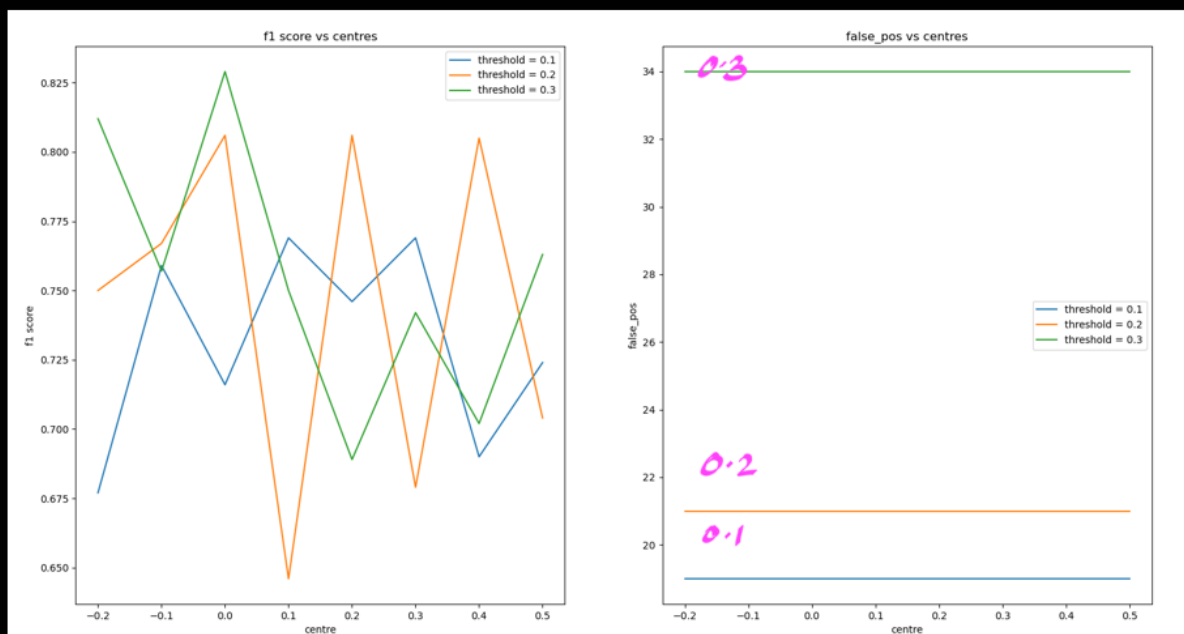
★ When WorkTimingsSatisfaction is dropped



→ Accuracy Vs false-negative (dropped)



→ Accuracy Vs False-positive (dropped)

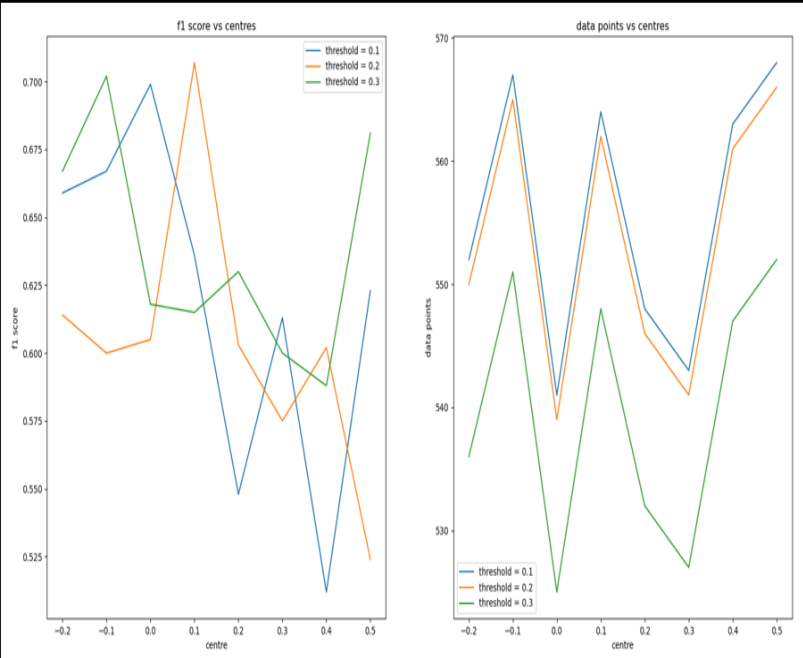


## → Accuracy & Related Data

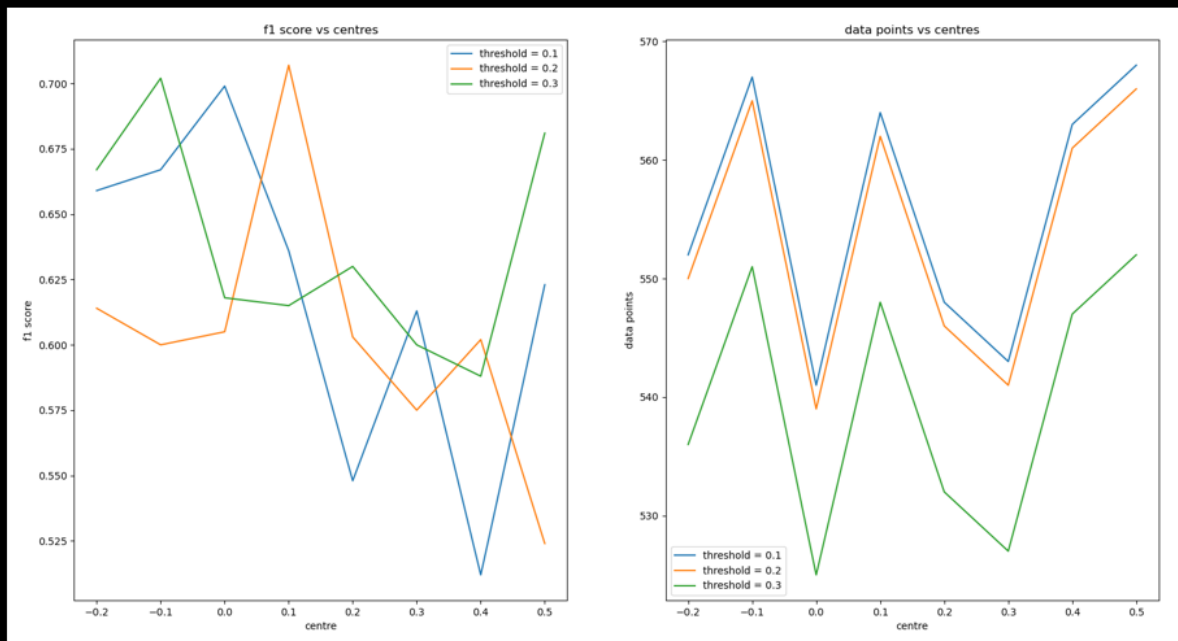
```
feature importance [0.0, 0.055, 0.0, 0.0, 0.0, 0.0, 0.0, 0.945]
feature importance [0.0, 0.073, 0.0, 0.0, 0.0, 0.0, 0.0, 0.927]
feature importance [0.0, 0.081, 0.0, 0.0, 0.0, 0.0, 0.0, 0.919]
data points [552, 567, 541, 564, 548, 543, 563, 568, 550, 565, 539, 562, 546, 541, 561, 566, 536, 551, 525, 548, 532, 527, 547, 552]
f1 scores [0.677, 0.759, 0.716, 0.769, 0.746, 0.769, 0.69, 0.724, 0.75, 0.767, 0.806, 0.646, 0.806, 0.679, 0.805, 0.704, 0.812, 0.757, 0.829, 0.75, 0.689, 0.742, 0.702, 0.763]
neutral [66, 51, 75, 53, 69, 74, 54, 48, 66, 51, 75, 53, 69, 74, 54, 48, 66, 51, 75, 53, 69, 74, 54, 48]
false negative [7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8]
false positive [19, 19, 19, 19, 19, 19, 19, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21]
f1_score_list [0.677 0.759 0.716 0.769 0.746 0.769 0.69 0.724]
false_pos_list [19 19 19 19 19 19 19 19]
false_neg_list [7 7 7 7 7 7 7]
neutral_list [66 51 75 53 69 74 54 48]
data_points_list [552 567 541 564 548 543 563 568]
centres_list [-0.2 -0.1 0. 0.1 0.2 0.3 0.4 0.5]
df f1_scores when WorkTimingsatisfaction is Not considered
```

	f1_scores	centres	thresholds	false_pos	false_neg	neutral	data_points
18	0.829	0.0	0.3	34	8	75	525
16	0.812	-0.2	0.3	34	8	66	536
12	0.806	0.2	0.2	21	7	69	546
10	0.806	0.0	0.2	21	7	75	539
14	0.805	0.4	0.2	21	7	54	561
3	0.769	0.1	0.1	19	7	53	564
5	0.769	0.3	0.1	19	7	74	543
9	0.767	-0.1	0.2	21	7	51	565
23	0.763	0.5	0.3	34	8	48	552
1	0.759	-0.1	0.1	19	7	51	567
17	0.757	-0.1	0.3	34	8	51	551
8	0.750	-0.2	0.2	21	7	66	550
19	0.750	0.1	0.3	34	8	53	548
4	0.746	0.2	0.1	19	7	69	548
21	0.742	0.3	0.3	34	8	74	527
7	0.724	0.5	0.1	19	7	48	568
2	0.716	0.0	0.1	19	7	75	541
15	0.704	0.5	0.2	21	7	48	566
22	0.702	0.4	0.3	34	8	54	547
6	0.690	0.4	0.1	19	7	54	563
20	0.689	0.2	0.3	34	8	69	532
13	0.679	0.3	0.2	21	7	74	541
0	0.677	-0.2	0.1	19	7	66	552
11	0.646	0.1	0.2	21	7	53	562

## \* Overtime & Work Timing Satisfaction dropped



## \* Overtime & Work Timings Satisfaction dropped



### Conclusion

What are your main conclusions?

The most important factor standing out as the reason for attrition was 'Work Timing Satisfaction' for the employees and followed by how often they are compelled to do Overtime.

Since the data was limited and we feel that only employees looking for some changes are visiting these sites like glass door and ambition box, so there is always a natural Bias within the data.

We tried to counter that by leveraging the fact that we had handwritten reviews by the employees, and we looked how subjective each review and did further data filtering so as to centre the data at appropriate centre for the neutrality of each review and then removed those data points.

We used this filtered data into our analysis for building a classifier to predict the risks of a employee attrition and gave the most prominent reason for it. The best classifier we got gave a f1 score of 0.9247.



**Evaluation criteria (max 30)**

**You have 12 minutes for your presentation.**

<b>Criterion</b>	<b>Marks</b>
Definition of scope, clarity of the presentation of scope <i>(Did you describe the problem clearly? Did you define the scope and goal?)</i>	5
Solution methodology <i>(Did you describe your proposed approach clearly?)</i>	5
Description of results <i>(Did you describe your results crisply, clearly, and completely? Did you meet the goals you set for your team?)</i>	5
Innovation, new insights, discussion <i>(Did you do something new? Did you create an 'Aha' moment for your audience?)</i>	5
Response to questions <i>(How good were your responses to the questions)</i>	5
Effort, substantiality, and technical depth <i>(How much did you achieve? Was it a hotch-potch, previous night-out effort, or did you plan and systematically make progress in your project (as evidenced by your presentation)?)</i>	5