Raise Prediction from IBM HR Analytics Employee Attrition and Performance Sample Dataset

Ali Babayev, Yusuf Samsum, Cagla Sozen

September 11, 2018

1 Introduction

This project aims to investigate the relationship between the attributes given in the *IBM HR Analytics Employee Attrition and Performance Sample Datasets.* Using *pandas, seaborn* and *matplotlib* libraries with Python, the project hopes to be able to classify the **PercentSalaryHike** with the chosen attributes by the end of Data Exploration according to their individual effects into three categories designated by the authors. Three classes defined are as the following

- **Low: 11-15**
- **Medium:16-20**
- **High:21-25**

2 Explaining the Dataset

2.1 The Data

The dataset consists of 1470 rows of fictional data prepared by IBM Data Scientists for utilizing Machine Learning in order to predict *Attrition.* However in the scope of this project, Machine Learning will be utilized in the same manner but the target value will be *PercentSalaryHike* and its values vary between the values 11-25. For increasing the accuracy in the first place, the target data will be classified into three as Low(L),Middle(M),High(H) with the values given in the **Introduction** section. All data will be used for visualization and understanding the relations, but the attributes with no decent relationship with PercentSalaryHike will not be included in the *Prediction* stage.

2.2 Attributes

There are 35 columns in the original dataset, three of them, *Employee Number, Over18* and *PercentSalaryHike*, will be discarded for the sake of this project since Percent Salary Hike is the target value and Attrition is a set of experimental values. Attributes can be listed as the following:

• Age	Age of the employee
• Attrition	Did the employee decide to leave the company?
• BusinessTravel	How often does the employee travel for the company?
• DailyRate	Daily salary level of the employee
• Department	Department of the company that the employee is currently working in
• DistanceFromHome	How far is the home of the employee from the company?
• Education	Level of education of the employee based o graduate schools
• EducationField	Graduation Department

- EmployeeCount How many employees does the employee work with?
- EmployeeNumber ID od the employee
- EnvironmentSatisfaction How satisfied is the employee from the company environment?
- Gender gender of the employee
- HourlyRate Monthly salary level of the employee
- JobInvolvement How involved is the employee with his/her job?
- JobLevel Level of the joc the emloyee is assigned
- JobRole What is the employee working as within the job?
- JobSatisfaction How satisfied is the employee with his/her job?
- MaritalStatus Is the employee married?
- MonthlyIncome Monthly salary of the employee
- MonthlyRate Monthly salary rate of the employee
- NumCompaniesWorked How many companies did the employee work with in the past?
- Over18 Is the employee over 18 years old?
- OverTime Does the employee work overtime?
- PercentSalaryHike Percentage of raise that the employee will get
- PerformanceRating How does the employee perform?
- RelationshipSatisfaction Is the employee satisfied with his/her relationship?
- StandardHours Standard working hours of the employee
- StockOptionLevel Stock options of the employee
- TotalWorkingYears Experience of the employee
- TrainingTimesLastYear How many times was the employee trained last year?
- WorkLifeBalance Time spent between work and outside?
- YearsAtCompany How long has the employee been working in the company?
- YearsInCurrentRole How long has the employee been working in the current position?
- YearsSinceLastPromotion How many years passed since the last promotion of the employee
- YearsWithCurrManager hfill How long has the employee been working with the current manager?

2.3 Data Exploration

In the first step of Data Exploration, simple data visualization techniques were used to understand the direct relationships between attributes and the target value.

Age

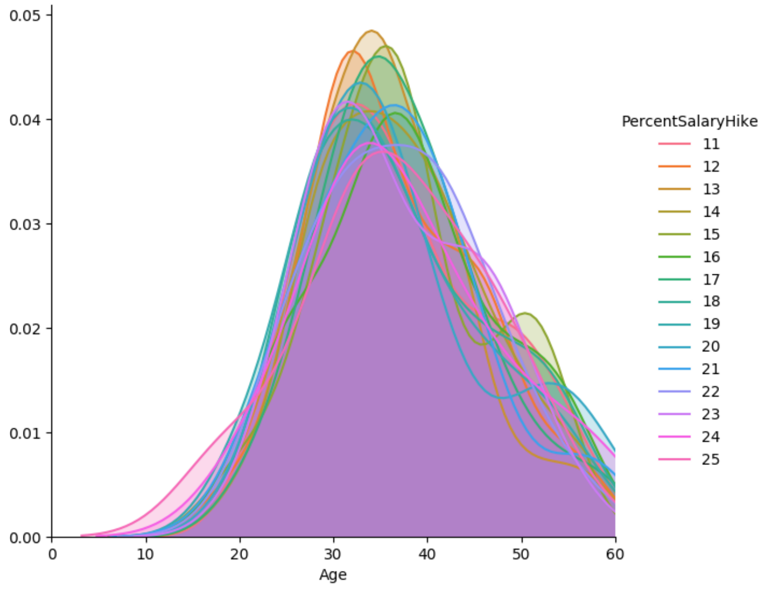

Figure 1: Age vs. Percent Salary Hike Graph

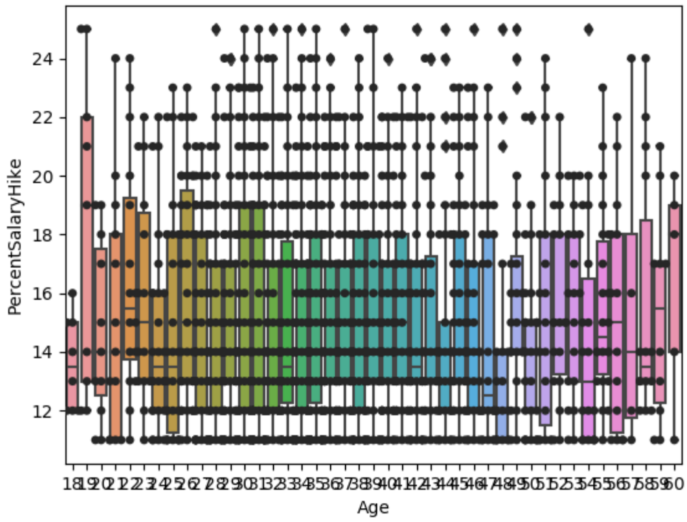

Figure 2: Age vs. Percent Salary Hike Box Graph

Figures 1 and **??** show the relationship between the age of an employee and percent salary hike. From the plot, it can be observed that although there isn't a significant distinction for an age range with higher percent salary hike, the data can be approximately fit onto a model.

Attrition

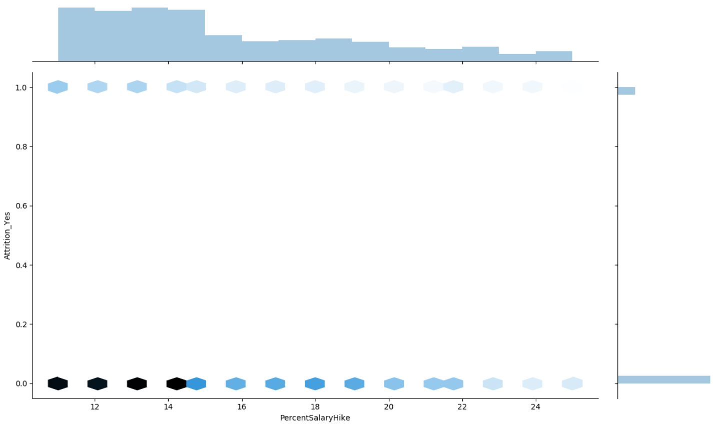

Figure 3: Attrition vs. Percent Salary Hike Graph

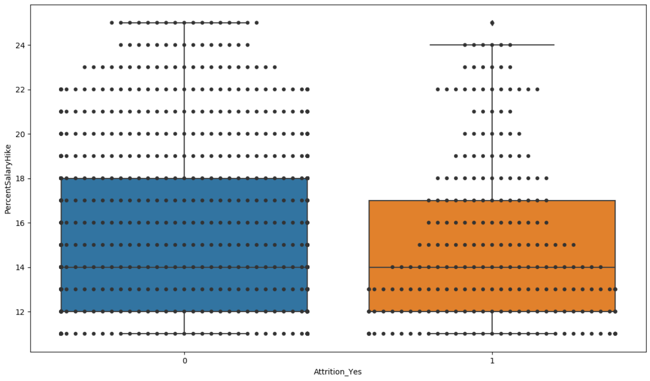

Figure 4: Attrition vs. Percent Salary Hike Box Graph

Figures 3 and **??** show the relationship attrition decision of an employee (Yes = 1, No = 0) and percent salary hike. It is observed in the graph that employees that did not decide for attrition are more likely to get a higher Percent Salary Hike.

Business Travel

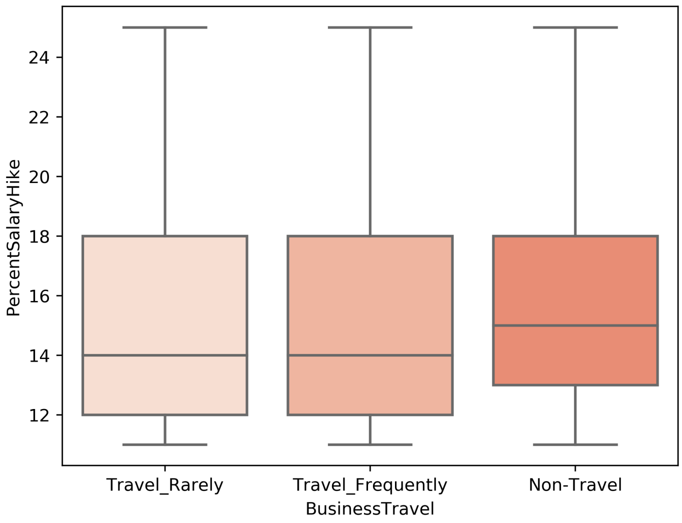

Figure 5: Business Travel vs. Percent Salary Hike Graph

Figure 5 shows the relationship between number of times an employee has traveled for the company and percent salary hike. By looking at the BoxPlot, it can be seen that Non-Travelers differentiate from Travel Rarely and Travel Frequently groups by having a median value of 15 and not having its lower quartile on higher values. But Travel Rarely and Travel Frequently are not visibly distinguishable.

Daily Rate

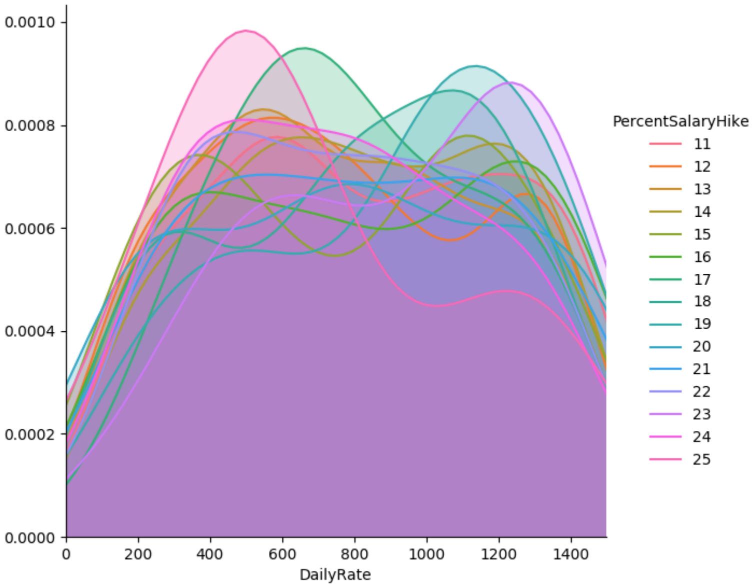

Figure 6: Daily Rate vs. Percent Salary Hike Graph

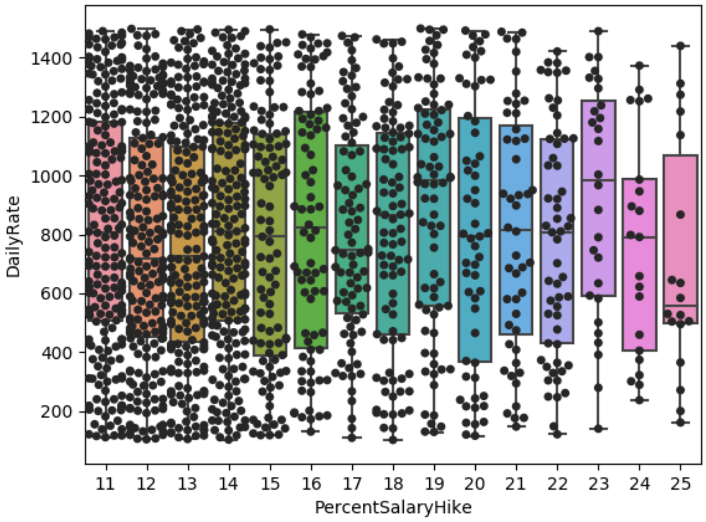

Figure 7: Daily Rate vs. Percent Salary Hike Box Graph

Figures 6 and ?? show the relationship between the daily rate of an employee and percent salary hike. The plot shows that Daily Rate of an employee is expected to effect the percent salary hike in a proportional manner. Especially in Figure ??, it is explicitly observed that employees with a daily rate higher than 600 are likely to get a Percent Salary Hike of 25.

Department

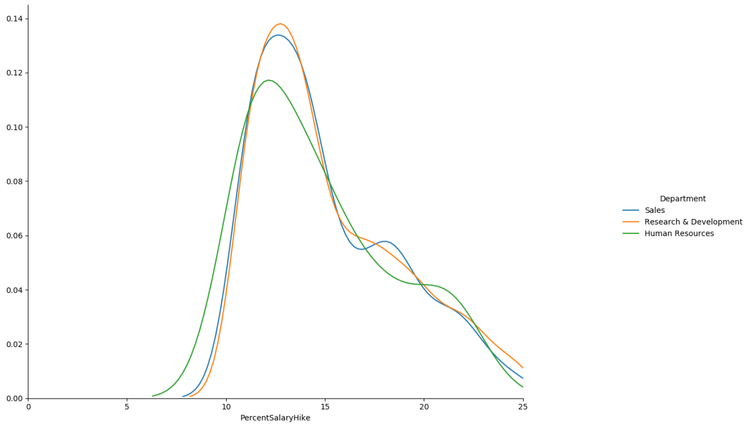

Figure 8: Department vs. Percent Salary Hike Graph

Figure 8 shows the relationship between the department of the employee and percent salary hike. Observation made from the plot is that Human Resources Department is less likely to get a higher Percent Salary Hike while Research and Development is likely to get a higher Percent Salary Hike, and Sales Department is in between the two.

Distance From Home

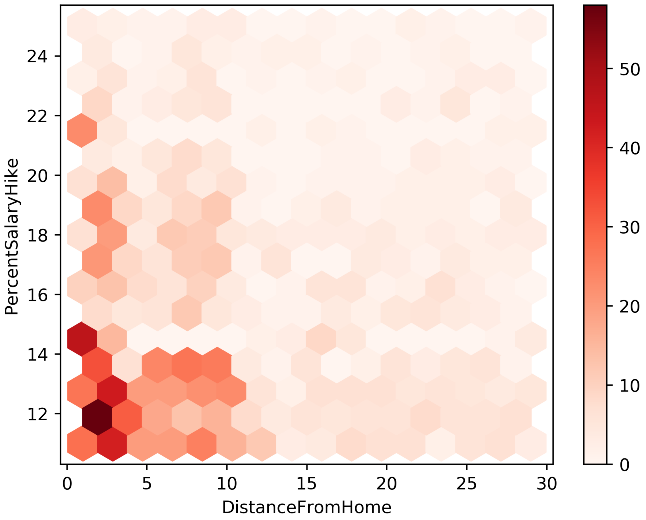

Figure 9: Distance From Home vs. Percent Salary Hike Graph

Figure 9 shows the relationship between the distance of the home of the employee from the company and percent salary hike. However the only visible information that can be deduced from this graph is that most employees who live closer to the company are in LOW salary hike bands.

Education Level

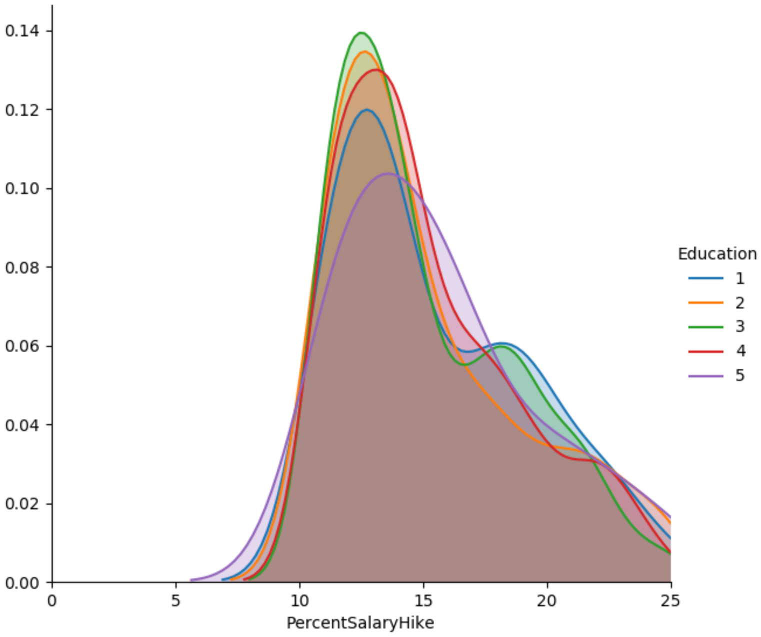

Figure 10: Education Level vs. Percent Salary Hike Graph

Figure **??** shows the relationship between the education level of the employee and percent salary hike. From the plot, it can be expected that employees with a education level higher than 3 is likely to get a higher Percent Salary Hike.

Education Field

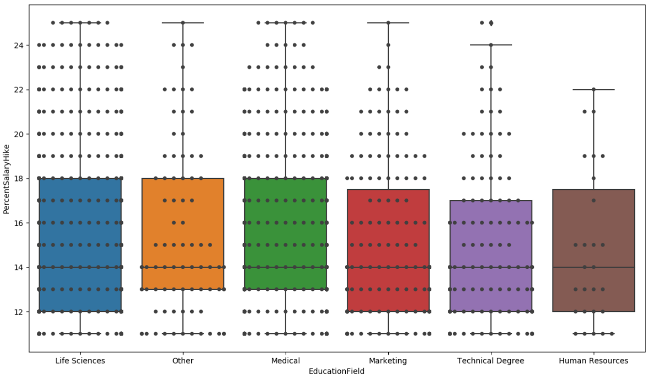

Figure 11: Education Field vs. Percent Salary Hike Graph

Figure **??** shows the relationship between the education field of the employee and percent salary hike.The boxplot shows the Percent salary Hike ranges for all 6 education fields where *Human Resources*, *Other* fields, *Medical* have employees accumulated within a higher Percent Salary Hikes. But for *Life Sciences*,*Medical* and *Technical* fields have outliers with the highest Percent Salary Hikes.

Employee Count

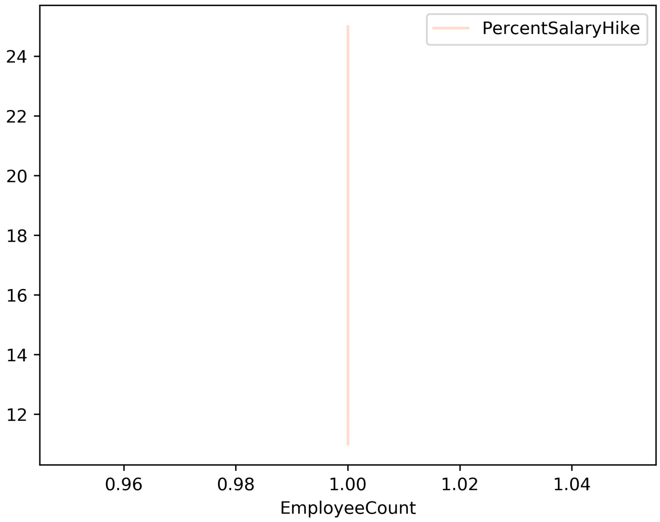

Figure 12: Employee Count vs. Percent Salary Hike Graph

Figure 12 shows the relationship between the employee count and Percent Salary Hike. However, here we discover that Employee Count is constant for all employees and do not effect percent salary hike

Environment Satisfaction

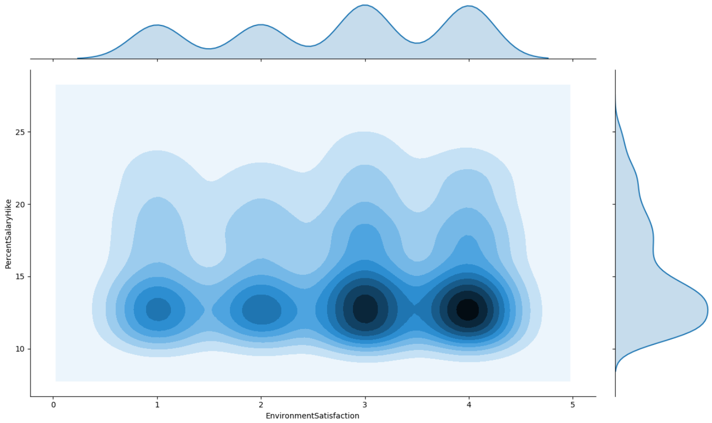

Figure 13: Environment Satisfaction vs. Percent Salary Hike Graph

Figure 13 shows the relationship between satisfaction of the employee within the environment and Percent Salary Hike. Here, it is explicitly seen that employees with 3-4 environment satisfaction is likely to get a percent salary hike of approximately 18. And the employees with an environment satisfaction of 2 is likely to get the lowest Percent Salary Hike.

Gender

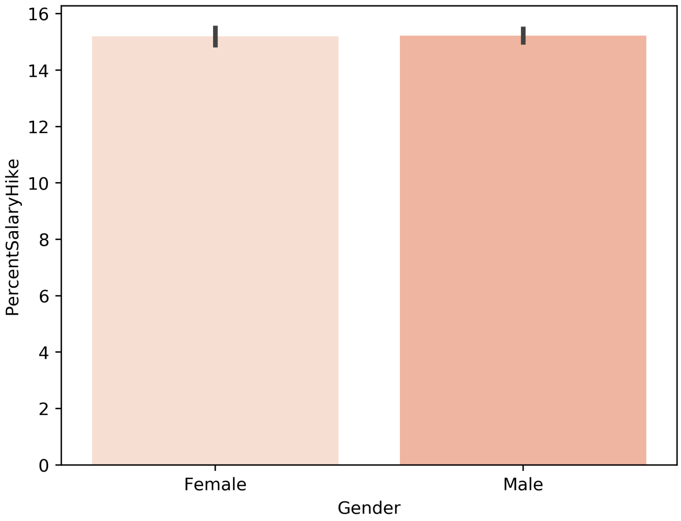

Figure 14: Gender vs. Percent Salary Hike Graph

Figure 14 shows the relationship between the gender of the employee and Percent Salary Hike. However, once again, here we discover that there is no correlation between Gender and Percent Salary Hike since the average of both genders have a raise around 15.

Hourly Rate

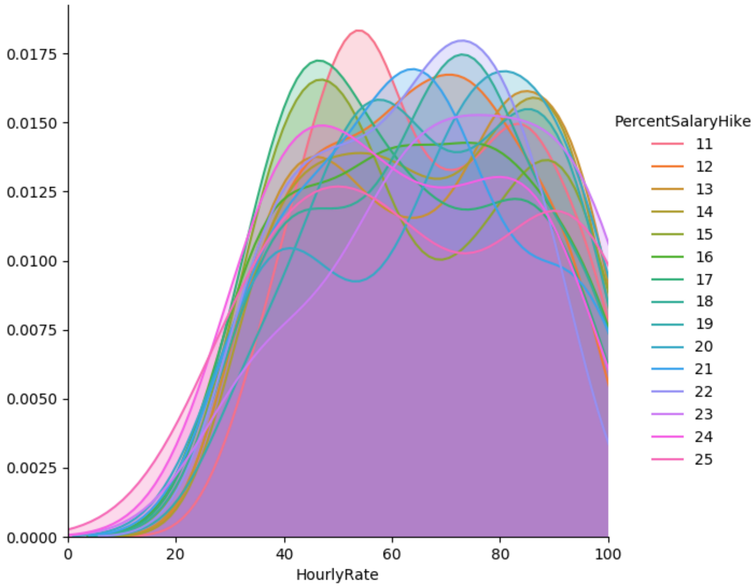

Figure 15: Hourly Rate vs. Percent Salary Hike Graph

Figure 15 shows the relationship between the Hourly Rate of the employee and Percent Salary Hike. From the plot,it can be observed that the Percent Salary Hike differs for all Hourly Rate values but the changes are not explicit.

Job Involvement

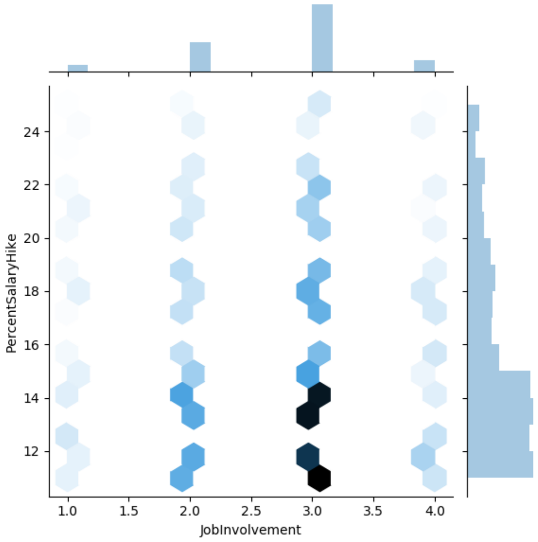

Figure 16: Job Involvement vs. Percent Salary Hike Graph

Figure 16 shows the relationship between the Job Involvement of the employee and Percent Salary Hike. It is explicitly observed that the employees with the highest Percent Salary Hike have a job Involvement of 3, but the increase in Percent Salary Hike is not proportional since for Job Involvements of 4, Percent Salary Hike decreases significantly.

Job Level

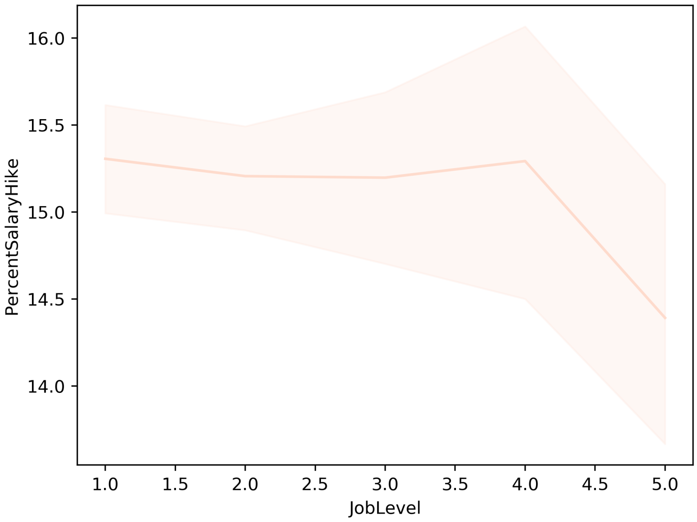

Figure 17: Job Level vs. Percent Salary Hike Graph

Figure 17 shows the relationship between job level of the employee and percent salary hike. The correlation we can observe from this graph is that average percent salary hike drops a point as Job Level variable goes above 4.0. For values lower than 4.0, percent salary hike is around 15.2, thus Job Level can be used as a variable for prediction.

Job Role

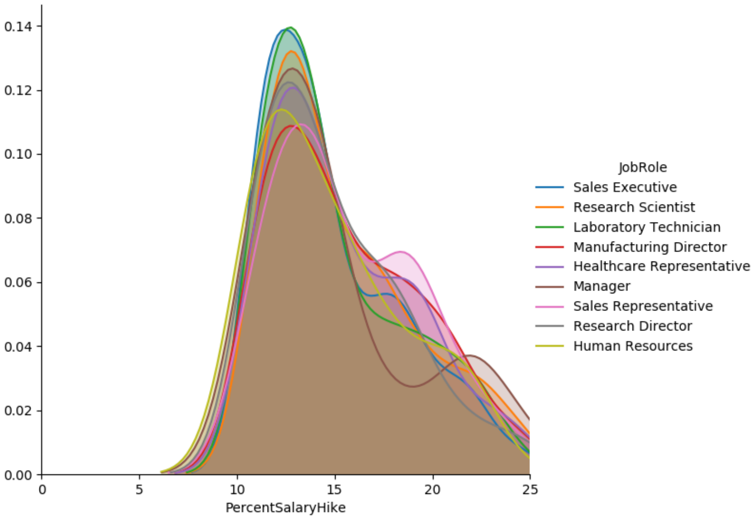

Figure 18: Job Role vs. Percent Salary Hike Graph

Figure 18 shows the relationship between job role of the employee and percent salary hike. From the plot, it can be deduced that Job Role is a parameter for Percent Salary Hike especially for Percent Salary Hikes above 15.

Job Satisfaction

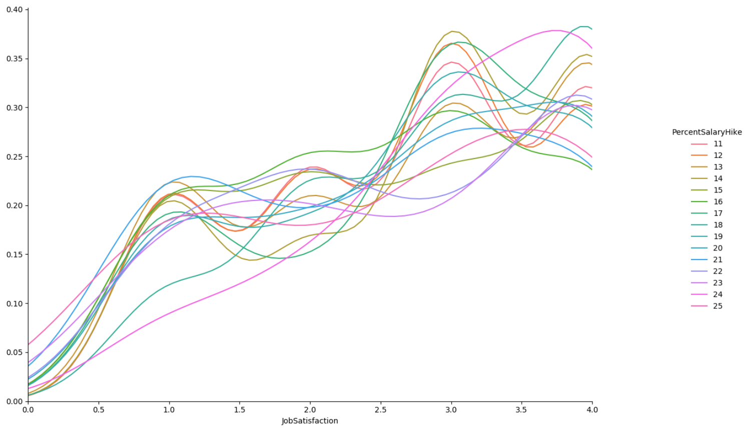

Figure 19: Job Satisfaction vs. Percent Salary Hike Graph

Figure 19 shows the relationship between job satisfaction of the employee and percent salary hike. In the plot, it can be observed that the highest Percent Salary Hikes peak around 3.5 Job Satisfaction while lower Percent Salary Hikes peak around 2.

Marital Status

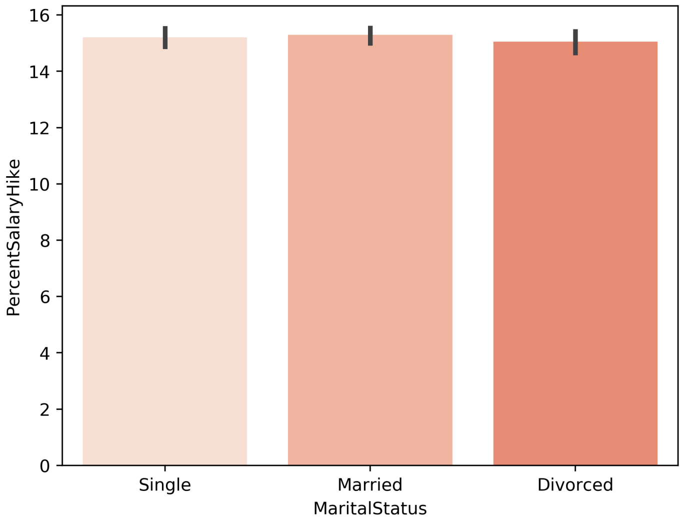

Figure 20: Marital Status vs. Percent Salary Hike Graph

Figure 20 shows the relationship between marital status of the employee and percent salary hike. Here, we can not observe any distinguishable difference between percent salary hikes other than a small drop for divorced employees, which does not help distinguish the employees.

Monthly Income

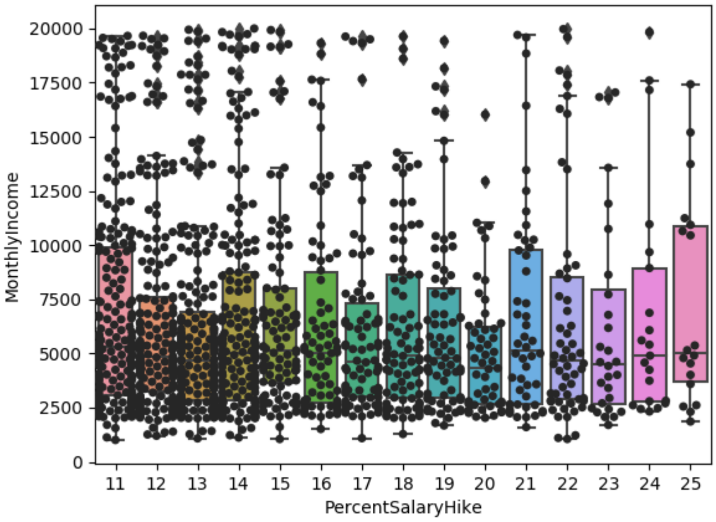

Figure 21: Monthly Income vs. Percent Salary Hike Graph

Figure 21 shows the relationship between the monthly income of the employee and percent salary hike. For this plot, no significant and traceable correlation was found for the sake of this project.

Monthly Rate

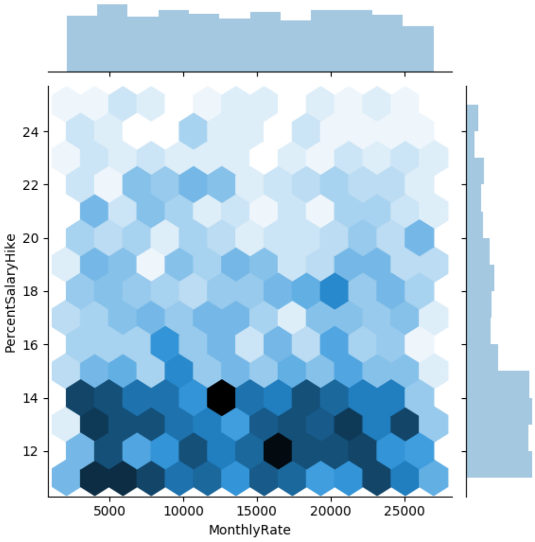

Figure 22: Monthly Rate vs. Percent Salary Hike Graph

Figure 22 shows the relationship between the montly rate of the employee and percent salary hike. Here, there was no meaningful correlation neither with Percent Salary Hike nor within Monthly Rate itself.

Number of Companies Worked

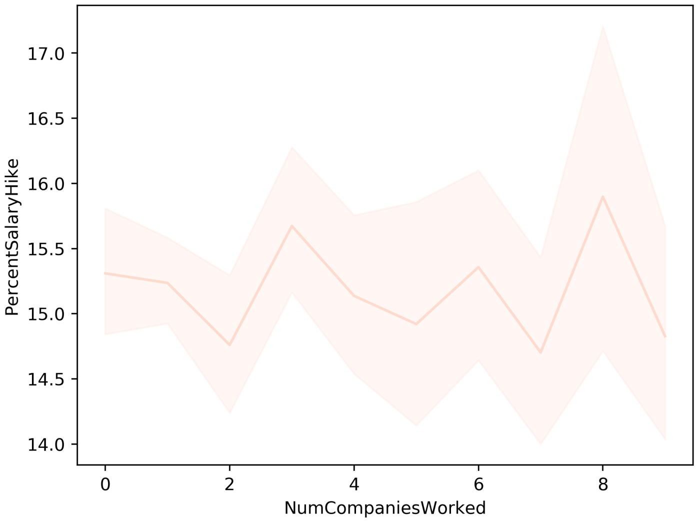

Figure 23: Number of Companies Worked vs. Percent Salary Hike Graph

Figure 23 shows the relationship between the number of companies worked by the employee and percent salary hike. Here, a proper correlation can not be observed however for all integer values between 0-8, there is continous function. So this attribute might be useful for predictions.

Overtime

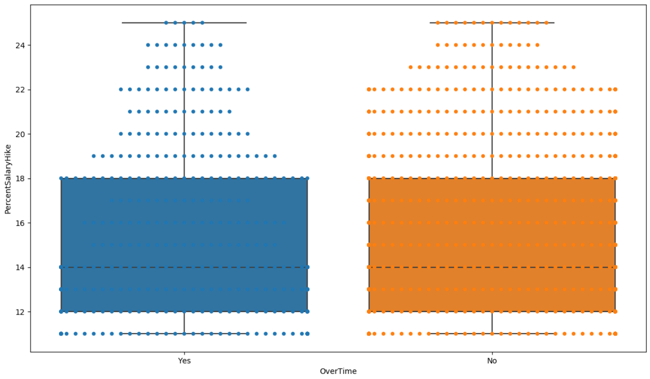

Figure 24: Overtime vs. Percent Salary Hike Graph

Figure 24 shows the relationship between the overtime working employees and percent salary hike. For this feature, none of the employees were distinguishable from the plot.

Performance Rating

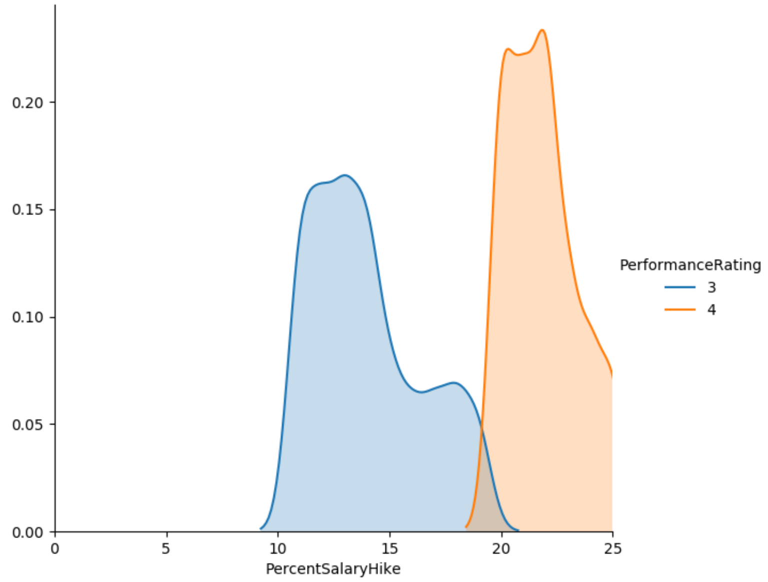

Figure 25: Performance Rating vs. Percent Salary Hike Graph

Figure 25 shows the relationship between the performance rating of the employees and percent salary hike. For this feature, there was an explicit distinction of employees with a performance rating of 3 and 4. For employees with a performance rating of 4, Percent Salary Hike was likely to be between 19-25 while for employees with a performance rating of 3, this range dropped to 9-21.

Relationship Satisfaction

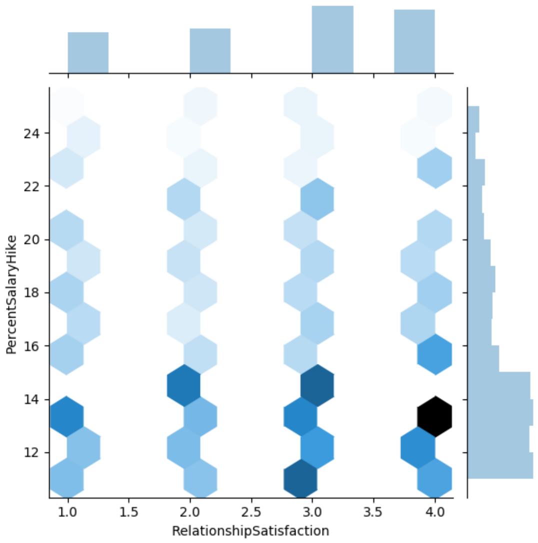

Figure 26: Working Hours vs. Percent Salary Hike Graph

Figure 26 shows the relationship between working hours and percent salary hike. In this plot, a little increase in Percent Salary Hike was observed for employees with a Relationship Satisfcation higher than 3.

Working Hours

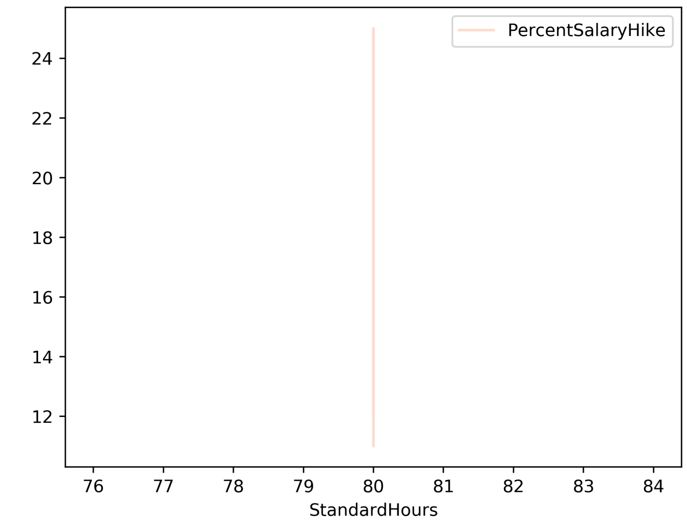

Figure 27: Working Hours vs. Percent Salary Hike Graph

Figure 27 shows the relationship between working hours and percent salary hike. Similar to employee count, working hours do not change for any employee in the company. Thus, standard hours is not a variable to use for predictions.

Stock Option Level

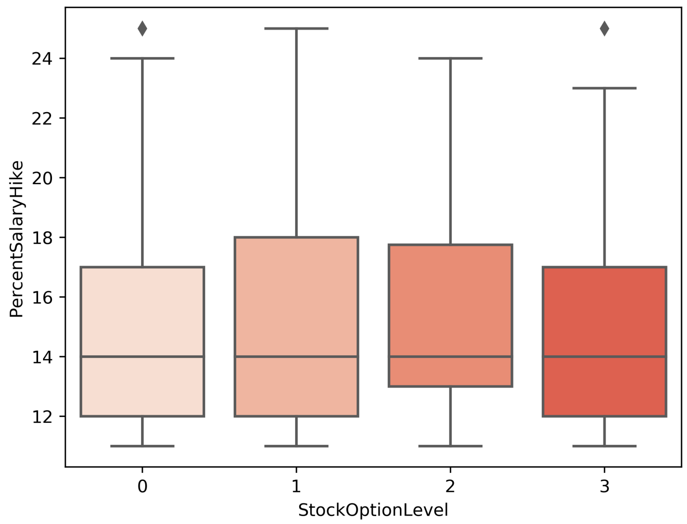

Figure 28: Stock Option Level vs. Percent Salary Hike Graph

Figure 28 shows the relationship between Stock Option Level and percent salary hike. As observed from the plot, although there are small differences in most of the higher quartiles and outliers, median of all Stock Option Level is on the same level. Thus, this feature will not explicitly distinguish an employee from another.

Total Working Years

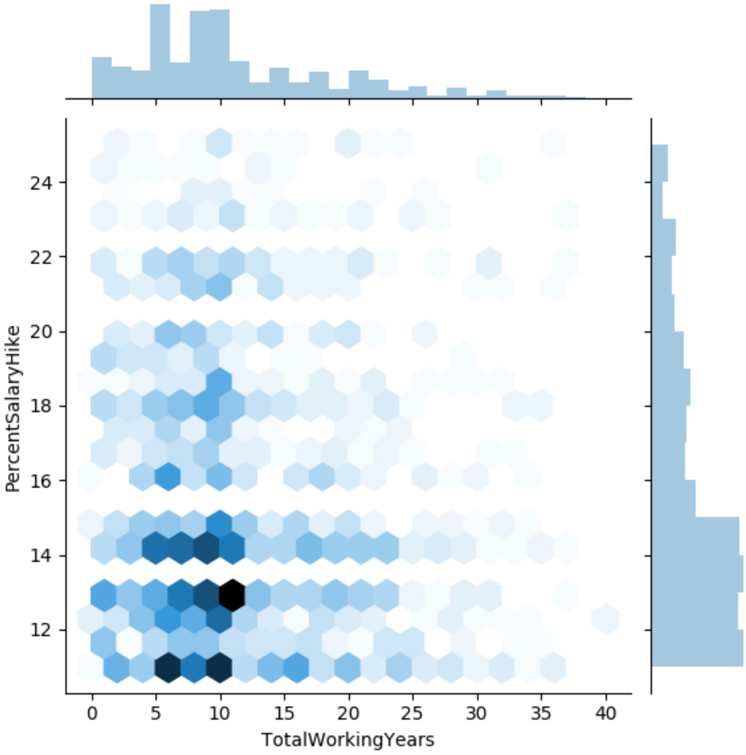

Figure 29: Total Working Years vs. Percent Salary Hike Graph

Figure 29 shows the relationship between the number of total working years of the employee and percent salary hike. Similar to the previous feature, no meaningful and distinguishing feature was observed.

Training Times Last Year

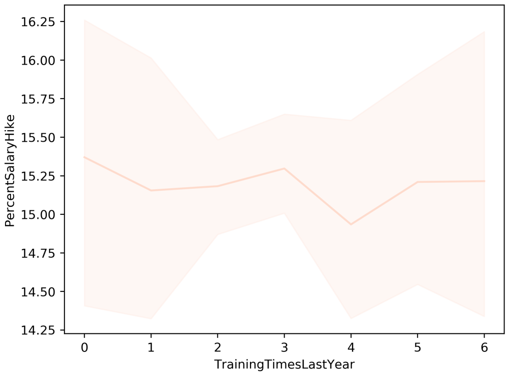

Figure 30: Training Times Last Year vs. Percent Salary Hike Graph

Figure 30 shows the relationship between the number of trainings the employee participated in last year and percent salary hike. Similar to the observation made for Number of Companies Worked, there is a function followed for each integer between 0-6. Hence,it can not be useful for predictions.

Work Life Balance

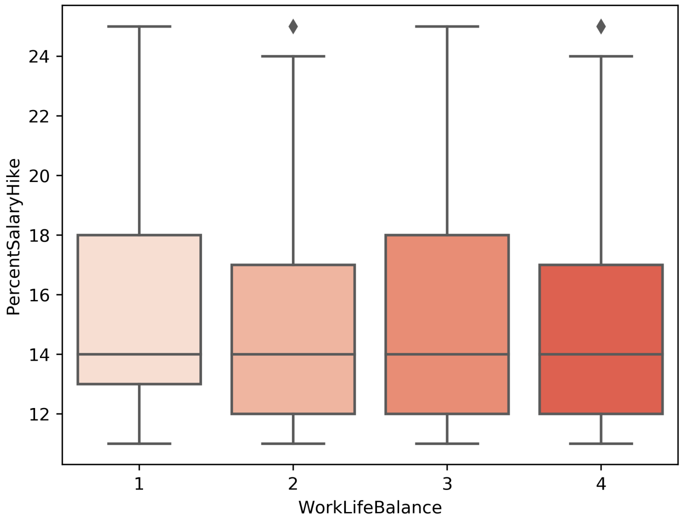

Figure 31: Work Life Balance vs. Percent Salary Hike Graph

Figure 31 shows the relationship between the work life balance of the employee and percent salary hike. In the plot, it was observed that there were slight differentiations for each value of Work Life Balance with medians on the same level.

Years At Company

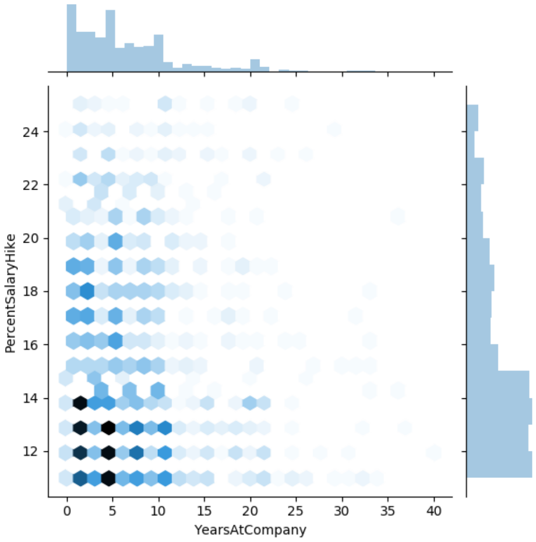

Figure 32: Years At Company vs. Percent Salary Hike Graph

Figure 32 shows the relationship between number of years the employee has been working in the company and percent salary hike. In the plot, it is explicitly observed that for some exact values of x, Percent Salary Hike was likely to be higher or lower.

Years In Current Role

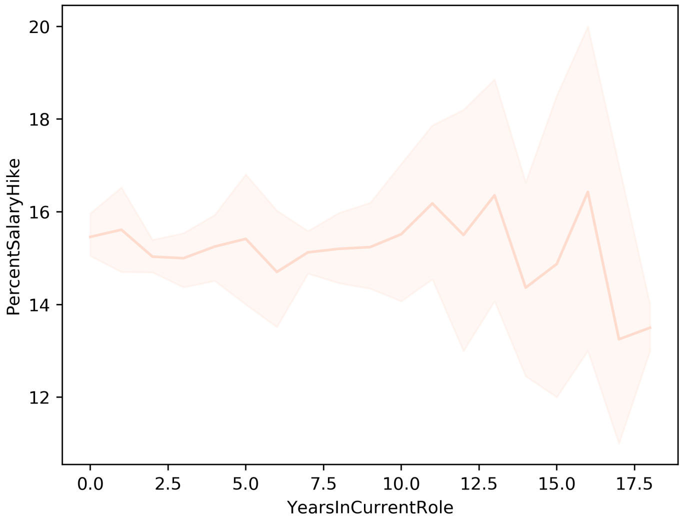

Figure 33: Years In Current Role vs. Percent Salary Hike Graph

Figure 33 shows the relationship between number of years the employee was in the current role and percent salary hike. The plot shows that for values of x, interval 0-10.0 may yield more accurate results and may be used for prediction.

Years Since Last Promotion

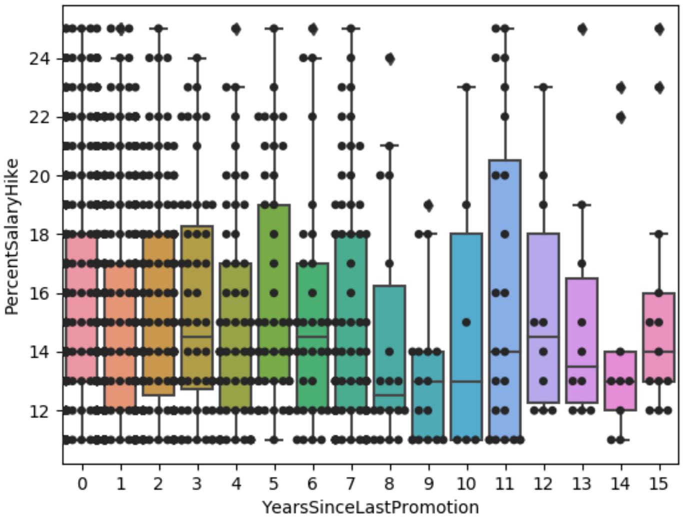

Figure 34: Years Since Last Promotion vs. Percent Salary Hike Graph

Figure 34 shows the relationship between the number of years since the employee was lastly promoted and percent salary hike. In the plot, although the correlation is not explicit, a decent pattern was implicitly likely to be followed.

Years With Current Manager

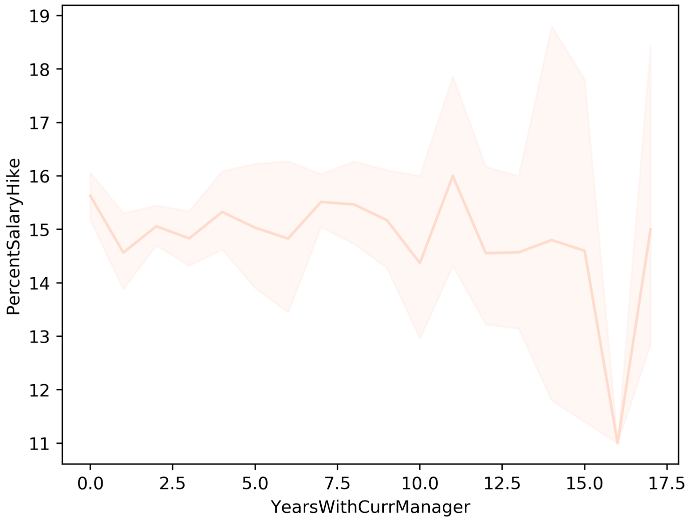

Figure 35: Years With Current Manager vs. Percent Salary Hike Graph

Figure 35 shows the relationship between number of years the employee was working with the current manager. This plot is similar to Figure 33 for the values around 0-8 however additionally, there is an explicit drop of values for employees with 16 years experience with the same manager.

3 Implementation

For the implementation of this project, the following environments and tools were used.

- **OS:** Linux
- **Language:** Python 2.7
- **IDE:** Pycharm
- **Libraries:** Pandas, Seaborn, matplotlib, sklearn

Implementation stage of the project began with implementing the visualizations given in the previous section. Code for the visualization stage can be found under the Python File *Visualization*.

After data was analysed in the light of the visualizations, relative features were selected to implement the prediction stage. A pandas DataFrame was used to hold the data in a manipulative and analyzable form within Python. The following features were selected in accordance with the data exploration,

- Age
- Attrition
- Business Travel
- Daily Rate
- Education Level
- Education Field
- Environment Satisfaction

- Hourly Rate
- Job Involvement
- Job Level
- Job Role
- Job Satisfaction
- Monthly Income
- Number of Companies Worked
- Stock Option Level
- Work Life Balance
- Years At Company
- Years In Current Role
- Years Since Last Promotion
- Years With Current Manager

Using the listed features, classification of Percent Salary Hikes into three categories was attempted. No pre-processing on the data was applied in the first stage. The following models were chosen to use for predictions were chosen respectively and used for classifications.

- LogisticRegression
- RandomForestClassifier
- SVC
- RidgeClassifier
- AdaBoostClassifier
- Perceptron

First, each model was evaluated using **Cross Value Scores** .

- **LogisticRegression:** 0.7367358035943018
- **RandomForestClassifier:** 0.7088746409653119
- **SVC:** 0.6251706425614152
- **RidgeClassifier:** 0.7414954023490579
- **AdaBoostClassifier:** 0.7414954023490579

Then, each model was evaluated separately for **Accuracy Scores** . Accuracy Scores and Classification Reports of each model can be found below.

3.1 Logistic Regression

Accuracy: 0.75

	precision	recall	f1-score	support
0	0.75	1.00	0.86	327
1	0.00	0.00	0.00	128
2	0.74	1.00	0.85	60
avg / total	0.56	0.75	0.64	515

3.2 Random Forest Classifier

Accuracy: 0.49

	precision	recall	f1-score	support
0	0.61	0.68	0.64	327
1	0.21	0.24	0.22	128
2	0.00	0.00	0.00	60
avg / total	0.44	0.49	0.46	515

3.3 SVC

Accuracy: 0.63

	precision	recall	f1-score	support
0	0.63	1.00	0.78	327
1	0.00	0.00	0.00	128
2	0.00	0.00	0.00	60
avg / total	0.40	0.63	0.49	515

3.4 Ridge Classifier

Accuracy: 0.63

	precision	recall	f1-score	support
0	0.63	1.00	0.78	327
1	0.00	0.00	0.00	128
2	0.00	0.00	0.00	60
avg / total	0.40	0.63	0.49	515

3.5 Ada Boost Classifier

Accuracy: 0.63

	precision	recall	f1-score	support
0	0.63	1.00	0.78	327
1	0.00	0.00	0.00	128
2	0.00	0.00	0.00	60
avg / total	0.40	0.63	0.49	515

As seen from the scores and reports, the best performing model for the first stage without any parameter adjustments and pre-processing was found to be Logistic Regression with an accuracy of %75.