Final Report: Statistical Modeling and Analysis Results for the Human Resources Case and Performance Modeling

Submitted to: GoalEarn

Report Prepared By: S. Hadis Hashemi Homami

March, 2023

1. Introduction

The main idea of a report is to analyze the data presented by the HR department. The purpose of this analysis is to make a model to determine the performance of employees, estimating employees on the verge of leaving the company, studying management performance, and finding the most suitable recruitment source for this particular company provided the data.
The remainder of this report is organized as follows. The section 2 describes why and how the data was preprocessed and the data extracted from the original data are introduced. The section 3 represents the investigation of a variety of keys which has an effect on the performance score given to each employee. In section 4 the Salary distribution is discussed. The section 5 is a description of the study on the probability of relations between salary and performance score with the turnover rate. The section 6 is dedicated to analyzing the performance of managers based on the performance of their employees. In section 7 detailed analysis is applied on the data to determine the best recruitment source and at last in section 8 the model created for predicting the performance of an employee based on some key features, is introduced.

2. Data Preprocessing

2.1. Data description

The data presented by the HR department contained 317 rows each representing an employee's record in 31 columns. 42 rows were identical in every column and 36 duplicated rows for 18 employees with some missing values in different columns.

2.2. Duplicate and missing values

They were handled by sorting the data in order of the number of values every row contained. In order to keep as much data as we can, the duplicated rows for the same names which held less data were dropped out of the dataset. There were 3 rows of data without employee ids (EmpID) and some employees had the same employee id. Holding the employee’s names accountable, duplicate and null values in the EmpID column were replaced by a random number in a range of the rest of the employee's ids.

2.3. Extracting additional databases

Since we are to fill the missing data as accurately as possible, 8 different databases are created. They hold data about names, positions, leavers, managers, salaries and performances. Each database had missing and duplicate values and they were handled by seeking help from the rest of data and with the help of these databases, inaccurate and missing values in the original database were replaced.

2.4. Feature Engineering

Some of the data are irrelevant to the study and some are excessive since they can be found implemented in other columns. Some of the columns have both numeric values (IDs) and text values also. With a brief check to make sure none of the useful data is put aside, they are removed from the database.
The next step is to change date values to time intervals so they are easier to cast for the analysis and modeling. The first column is the date of birth or DOB in the database. As anticipated, there are missing values among them. Since they are not more than 0.7% of the data, they are filled with the mode of the data in the relative column. Although there was a challenge hidden in this column which was formatting differences between values, the column data got cleaned and replaced with the age of each employee. The second column is the hiring date which had the same challenges in formatting and the missing values were about 1% of the number of rows. The third column was the termination date and with the help for this column and the hire date column, the employment interval of each employee was derived.
There was a column named Termreason in the data and it was studied in order to know whether or not it holds useful data and a double check for the EmploymentStatus data in the process. After examination, the column was dropped.

2.5. Outliers

Next Step is to detect outliers in data to prevent them from meddling with the analysis. Since the data is about human resources, statistical Gaussian behavior is expected from data so the statistical approach is selected to fetch the outliers. In the age column, statistics found employees over 70 years old outliers. It is unlikely but possible. 2 more rows were detected as outliers which contained obvious wrong values. The wrong values were dropped. Same analysis was performed on the employment interval and no salient outliers were found. Visualization came to the rescue of finding outliers in the LastPerformanceReview_Date column. The wrong value was dropped.

Next, the missing values scattered through the data were handled logically. For example in the sex column, missing data were filled by using the employee's names. Missing values in the absence column filled with zeros and so on.

Finally, data is clean and ready for further processing.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 296 entries, 109 to 289
Data columns (total 22 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   EmpID                       296 non-null    float64
 1   PerfScoreID                 296 non-null    int64  
 2   Salary                      296 non-null    float64
 3   State                       296 non-null    object 
 4   Sex                         296 non-null    object 
 5   MaritalDesc                 296 non-null    object 
 6   CitizenDesc                 296 non-null    object 
 7   RaceDesc                    296 non-null    object 
 8   EmploymentStatus            296 non-null    object 
 9   Department                  296 non-null    object 
 10  RecruitmentSource           296 non-null    object 
 11  PerformanceScore            296 non-null    object 
 12  EngagementSurvey            296 non-null    float64
 13  EmpSatisfaction             296 non-null    int64  
 14  SpecialProjectsCount        296 non-null    int64  
 15  LastPerformanceReview_Date  296 non-null    object 
 16  DaysLateLast30              296 non-null    int64  
 17  Absences                    296 non-null    float64
 18  ManagerID                   296 non-null    float64
 19  PositionID                  295 non-null    float64
 20  age                         296 non-null    int64  
 21  emplmnt_intrvl              296 non-null    float64
dtypes: float64(7), int64(5), object(10)
memory usage: 53.2+ KB
```

Table 1. Preprocessed Data

2.6. Labeling and Discretizing data

In order to smoothing the path of the analysis, discretizing continuous data such as salary and assigning a numeric value to each class of the columns containing text values were necessary. “KBinsDiscretizer” or “EqualWidthDiscretiser” were tested on salary values, but discretizing them with manually showed better results. Other columns which should have changed to numeric data were: ‘EmploymentStatus’, ‘RecruitmentSource’, ‘State’, ‘Sex’, ‘MaritalDesc’, ‘CitizenDesc’ and ‘RaceDesc’ columns.

2.7. Extracting data divided by department names

Since data is complete and preprocessed, extracting data from employees working in each department is convenient. Each department has an excel file filled with its employees' data.
Now that the preprocessing part is fully covered, analysis can be started and questions can be answered.

3. Features Affect on Performance

Studying which and how any of the factors play a role in employees performance is necessary as always. Correlation is one of the easy yet strong ways to determine the relation between factors of the study.
Unfortunately the result of the correlation was inconclusive.

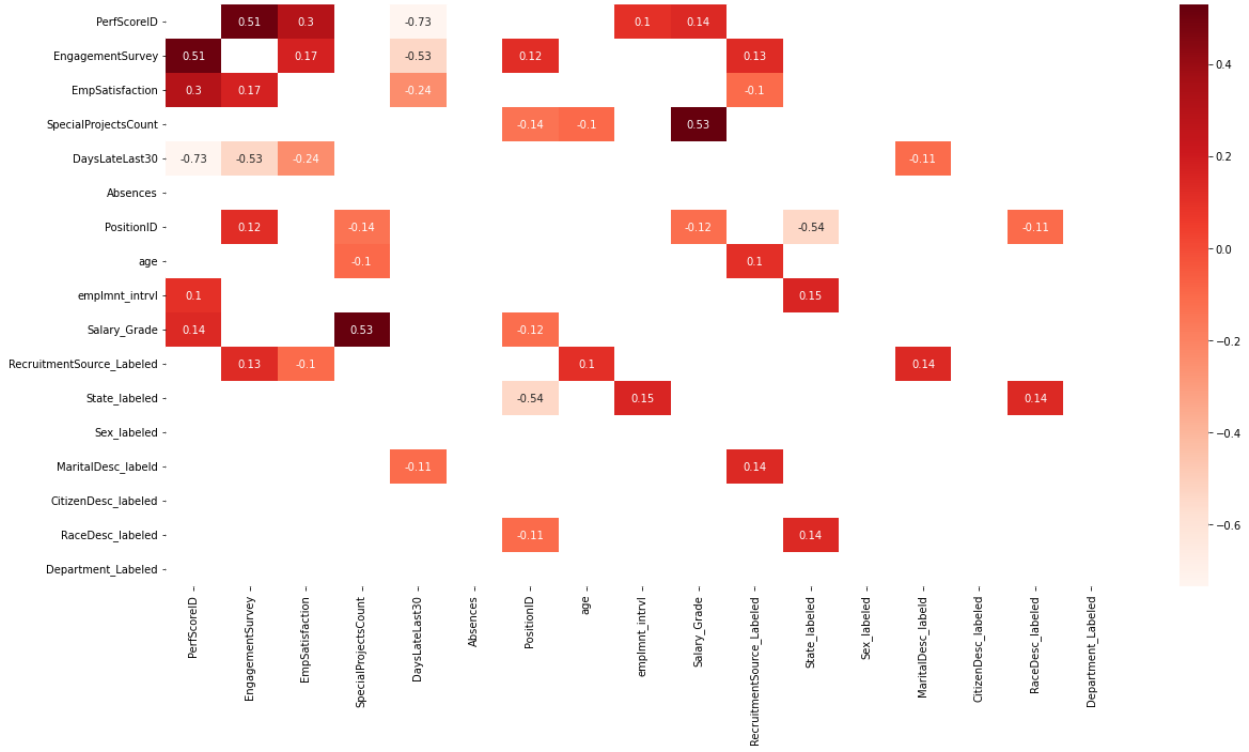

Chart 1. Correlation

Although it can be seen that ‘DaysLateLast30’ has a serious negative effect on performance scores, yet other factors are not strongly effective. "EngagementSurvey" and "EmpSatisfaction" have the most positive impact on performance scores as expected. "Salary" has a positive role in performance too although the effect is low. As statistics dictate other than ‘‘DaysLateLast30’ ‘ we should roll out every factor. But we know that is not quite right.

I personally believe factors mentioned above all have some part in performance score. Some are weaker than others. So features should get narrowed down enough to reach a conclusive result.

Diving into the data surrounding the performance score, results are as below:

1. Admin office and software engineering departments have the least number of employees with performance scores less than average. 20% of employees in the software engineering department have exceeded the expected performance. Production department, which has the most number of employees in the company, has only 11% of its employees, has failed to fully meet the expected performance.

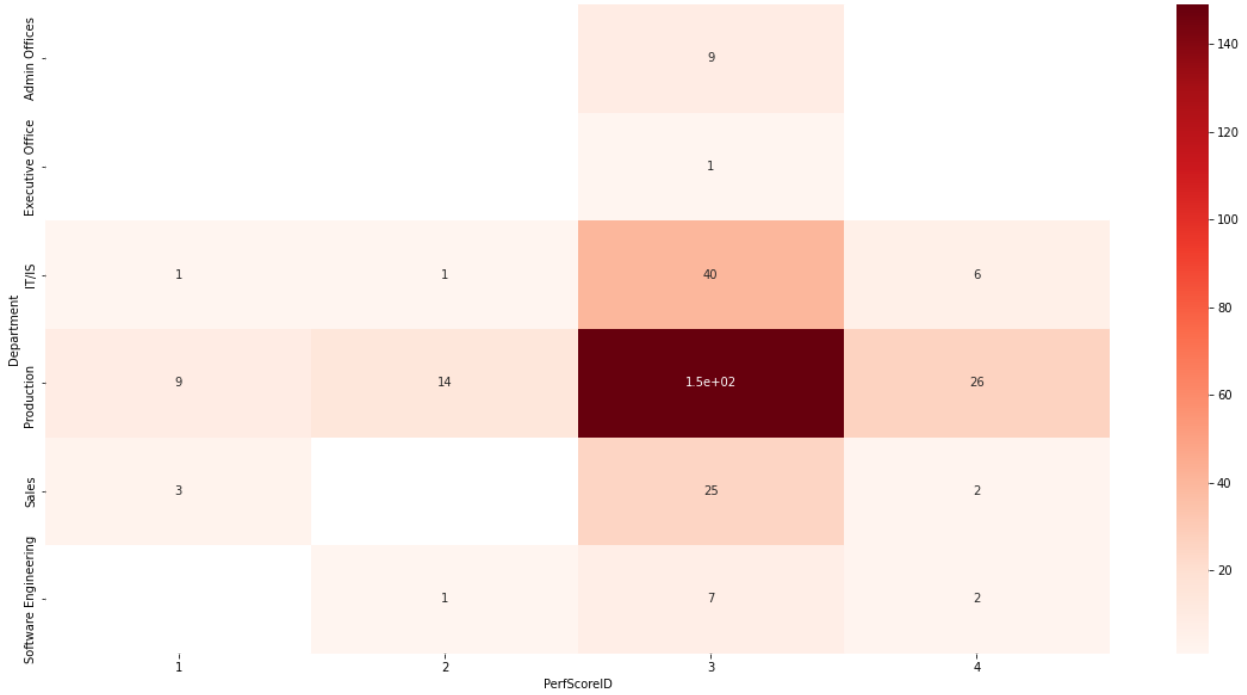

Chart 2. Departments/Performance Scores

2. Employees with higher engagement surveys have higher performance scores.
3. Employees with higher satisfaction have higher performance scores.
4. Employees with a higher number of special projects have higher performance scores.
5. 11% of area sales managers and 20% of network engineers have failed to meet expectations in performance.

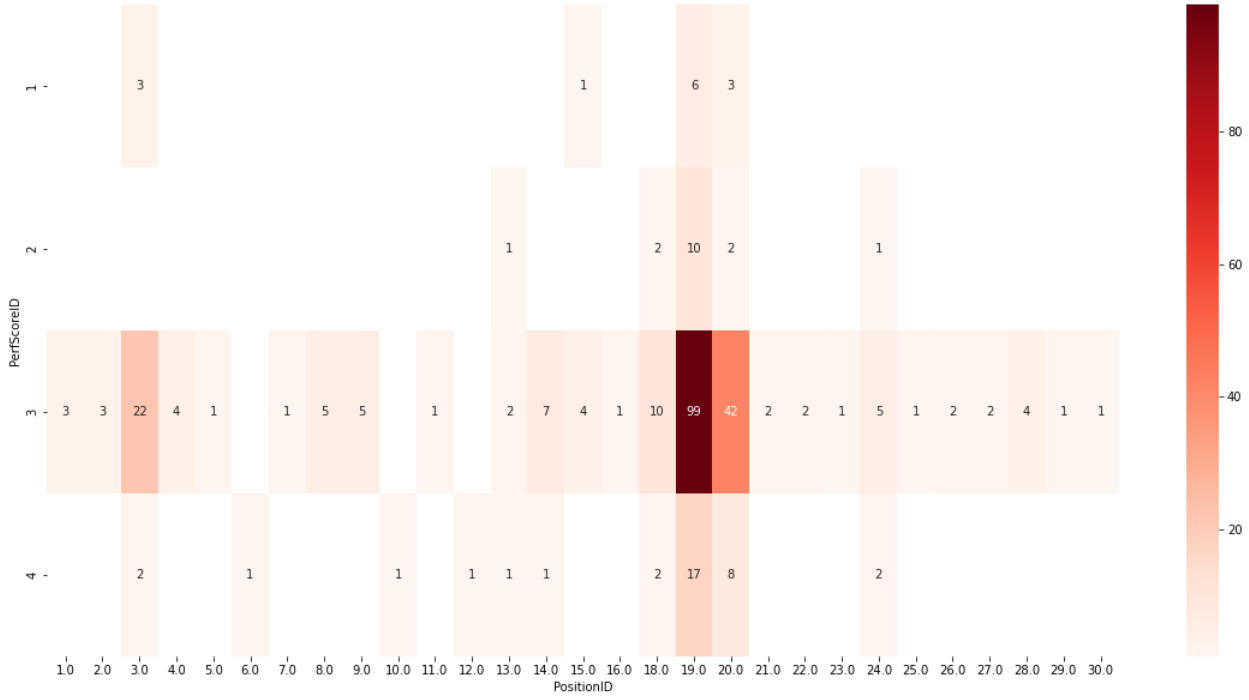

Chart 3. Positions / Performance Scores

6. Employees with age above 37 have better overall performance in the company, which is expected due to more work experience they have compared to their younger coworkers. Only 4% of employees with age above 50 have failed to meet the expected performance.

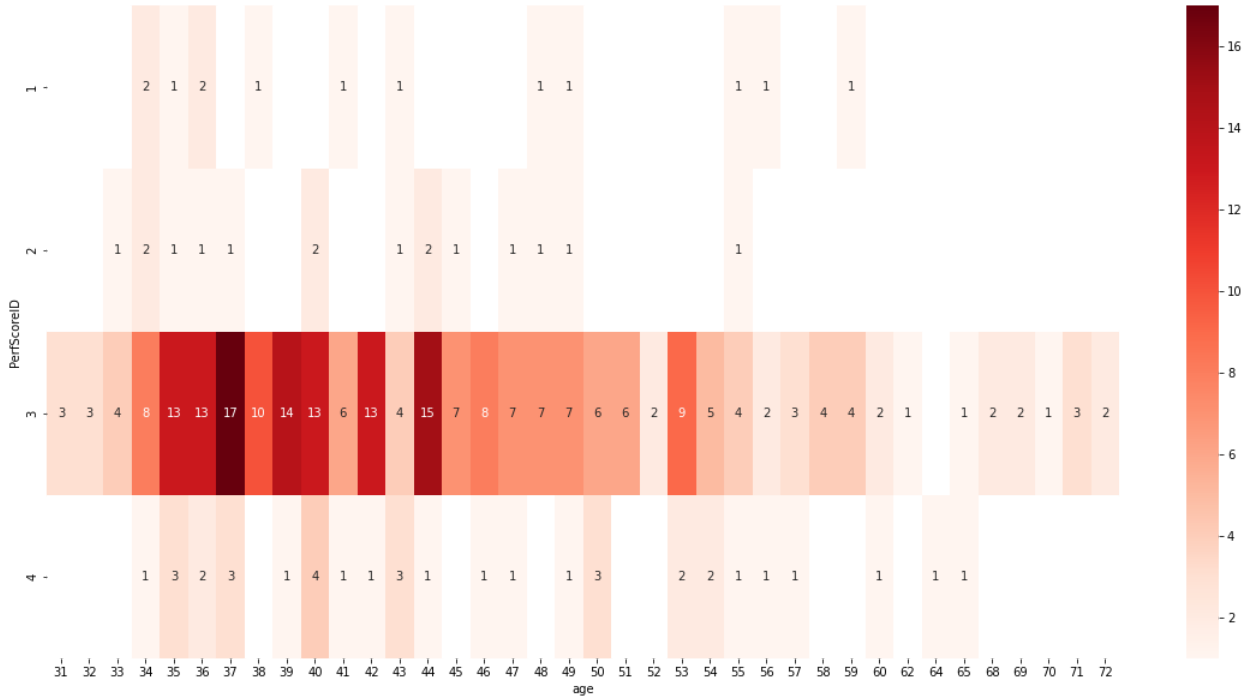

Chart 4. Age / Performance Scores

7. Vast majority of the employees have been working in the company between 6-11 years. The longer the employees stay in the company, the higher their performance goes.

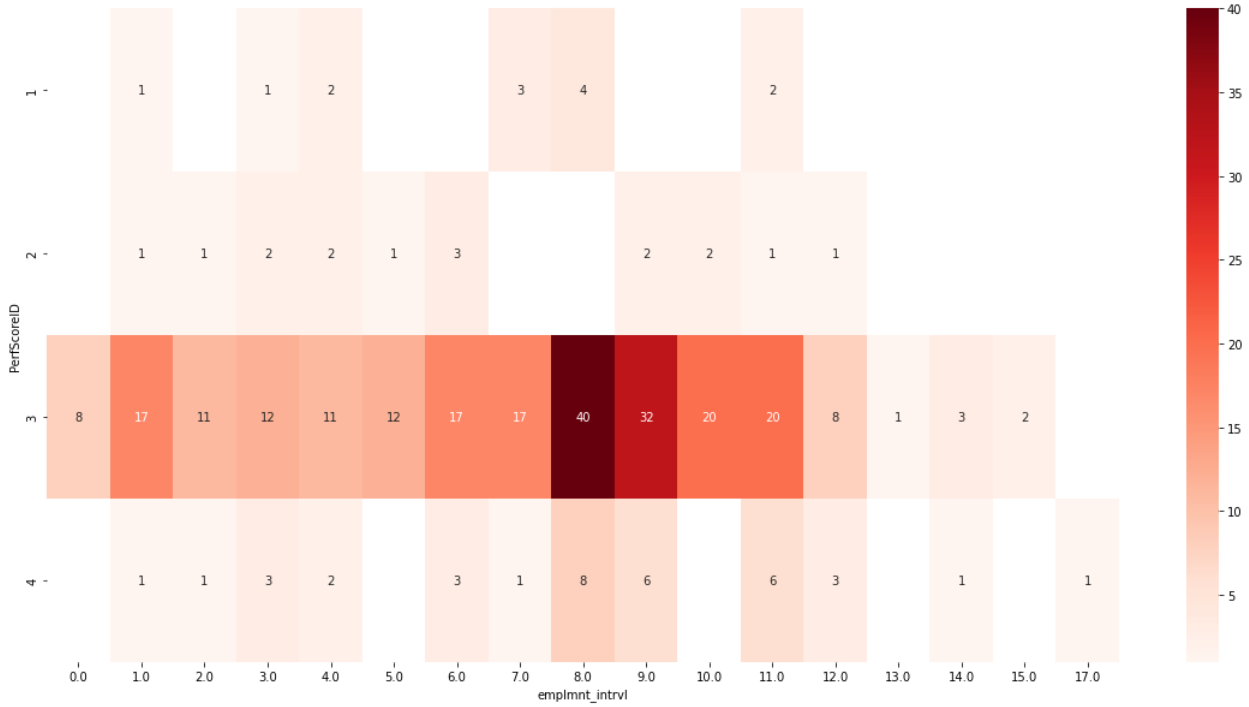

Chart 5. Employment Interval / Performance Score

8. The highest percentage of employees with a low performance score is 12% and their salaries are less than 70000. Obviously higher salaries have less performance failures.
9. Women are 58% of employees and there is no noticeable difference between performance of men and women in the company.
10. American Indian or Alaska Natives and Hispanic employees never had any failure in performance. 8% of white employees have failed the performance evaluation and as for the rest, about 12% of black or asians have performance score less than expected score.

4. Salary Distribution

Understanding how salary is distributed requires studying every factor separately. But first of all, it has to be checked whether or not the salary grading has been done correctly. By plotting salary against departments and salary grades against it, it is shown the gradings did not change the results, only simplified the analysis.

Right here, salary distribution between departments can be seen. The plots demonstrate the lowest average of salaries belongs to the production department and if the executive office has

been put aside, the highest average salaries goes to IT/IS and software engineering departments.

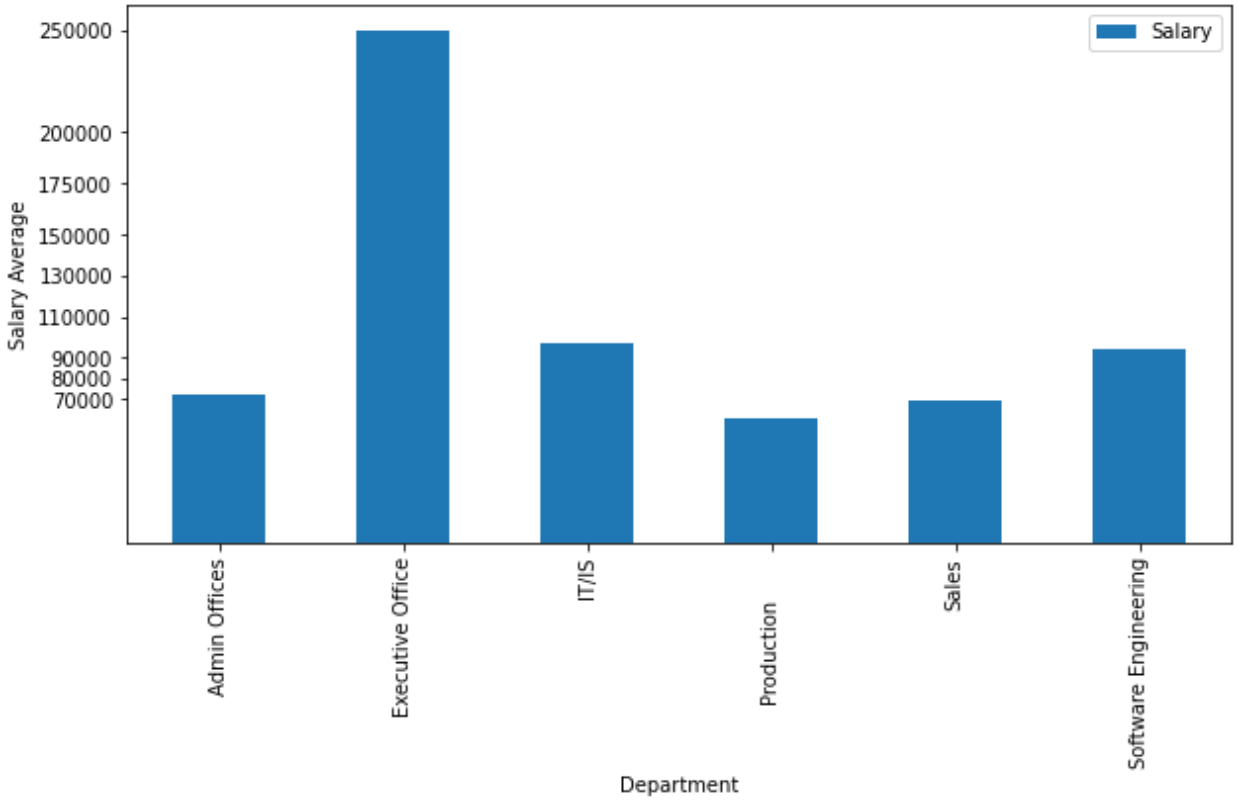

Chart 6. Salary Distribution / Departments

70% of employees receive less than 70000. After CEO and directors IT managers receive the highest salaries. Administrative assistants receive the least salaries and production technicians and accountants follow them.As expected and covered before, employees with higher performance scores receive higher salaries. MA employees receive slightly less salary compared to employees from the rest of the states. Male employees receive slightly more salary compared to female employees. US citizens among employees receive less salary compared to other kinds of work permitted employees. American Indians or Alaska Natives and Hispanics receive more salary compared to others.

5. Salary, Performance and Leavers

Conclusions derived from analyzing the behavior of turnover rate and salary or performance scores are listed below:

1. In the admin office the average salary of employees whose employment was terminated for cause, is noticeably less than active employees who left the company voluntarily. In IT/IS and software engineering departments the story is completely the opposite. In the production department active and terminated for cause leavers had similar salary average but voluntarily leavers had less salary average. In the sales department

however the story is different. Voluntarily Terminated employees received more salary compared to active employees and terminated for cause leavers received a lot less.

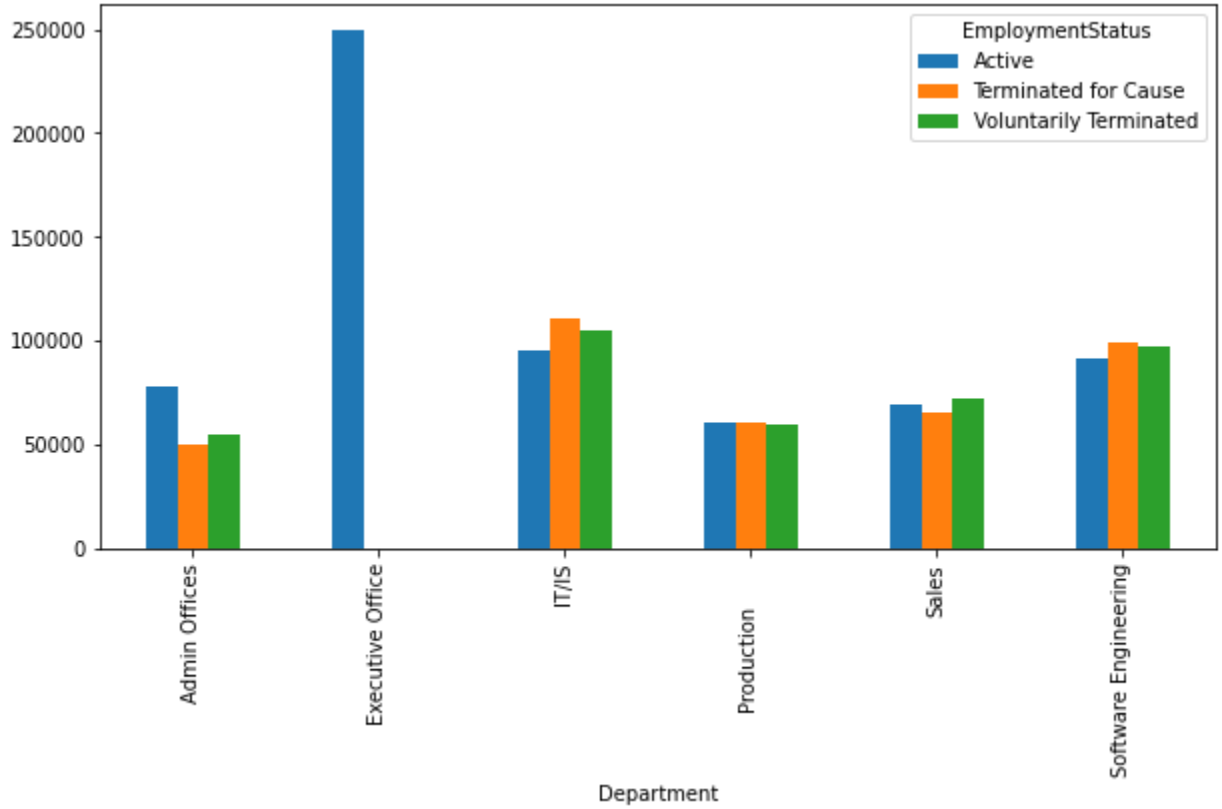

Chart 7. Salary / Department

2. In the admin office, leavers (voluntarily and terminated for cause) had the same performance scores as active employees. In IT/IS active employees have better performance than leavers. In software engineering and sales departments, the situation is different. Performance scores of voluntary leavers were higher than active employees. It looks like in the sales department, employees with higher performance scores were drawn to other companies.
3. The number of voluntary leavers in the production department is almost 7 times more than terminated for cause employees.In other departments it is nearly 2 times more.

In conclusion, making models for predicting whether or not an employee will leave the company can be done in each department separately but not in the company in general.

6. Manager Performance

Evaluation of managers' performance based on the performance of the employees who work under their supervision is required. Since there are combinations of performance scores among each manager's subordinates, creating a new score will help. The logic behind the new score is: (1) Expected outcome of performance evaluation is ‘Fully Meet’, so this performance score will

be the base of scoring. (2) Desired outcome of performance score is ‘Exceeds’, so this performance score will have higher score. (3) ‘Need Improvement’ is not the expected outcome but it has a weight of its own. (4) ‘PIP’ is the score we don't like to see, so it wont enter our calculations. As described, the formula for our new score is:
Managers performance score= (Count of ‘Exceeds’*2 + count of ‘Fully meet’ + count of ‘Needs Improvement’*.5 + count of ‘PIP’*0) / count of total subordinates
Managers with a score of 1 will be labeled as ‘Fully Meet’, managers with a score higher than 1 will be labeled as ‘Exceeds’ and managers with a score less than 1 will be labeled ‘Needs Improvement’.
After calculations, Jennifer Zamora, Eric Dougall and Alex Sweetwater are the top 3 managers based on their subordinates' performance. John Smith, Peter Monroe and Michael Albert won the least scores. It is worth mentioning that Peter Monroe and Michael Albert are no longer active employees.

7. Best Recruitment Source

Looking through data of hiring and recruitment, one can see the history of recruitment sources in the company. During this year, most of the employees were hired through Indeed. For the past 2 years google search provided most of the candidates for the hiring team. 3 years ago LinkedIn played a big role in hiring new employees. Of course during the 3 and 4 years ago there was a diversity job fair and it had a big part in recruitment also. 5 years ago the main recruitment source was Indeed. 6 and 7 years ago Indeed, LinkedIn and google search were used together and almost equally. 7 and 8 years ago employee referral was taken seriously and 8 and 9 years ago Indeed and LinkedIn again were the main source of recruitment.

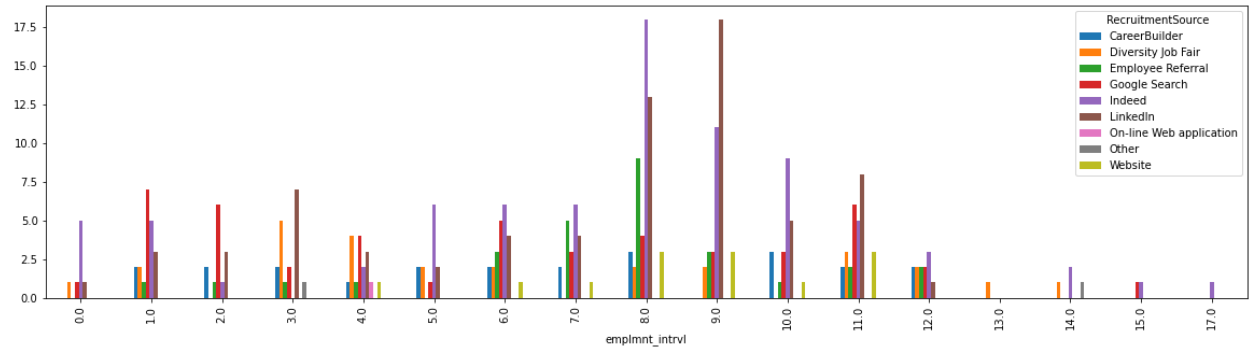

Chart 8. Performance Scores / Employment Interval / Recruitment Source

Similar to how the analysis in the past section was handled, a new score is introduced to evaluate the effectiveness of each recruitment source. First a new dataframe is formed with some of the features which can help during the process. The features were normalized and the result is as below:

RecruitmentSource	recref_score
Employee Referral	0.672150
On-line Web application	0.623529
Indeed	0.616921
Diversity Job Fair	0.609172
LinkedIn	0.599636
Other	0.585000
Google Search	0.577120
CareerBuilder	0.566701
Website	0.547534

Table 2. Recruitment Sources Scores

8. Performance Model

So it came down to modeling the performance. Here is the plan: The first question is which features are helping the model and which are not? RFE library helps the procedure and 8 features are selected for modelizing. The second question is which model suits the data best? There is a solution for that. One needs to test a lot of models to find the most suitable one. Fortunately there is a library for the rescue. Lazyprediction will test all classification models on the data in a few seconds and report back. Using the lazyprediction library, some models are selected. Next step is the hyperparameter tuning which is handled with the help of Gridsearchcv. Confusion matrix shows some of the models selected were overfit on the data and the best model is ExtraTreesClassifier. Now it is time to recreate the data which is going to be fed to the ExtraTreesClassifier. At last the model is done and scores are listed below:
Precision: 81%
Recall: 85%
F1 score: 81%
And the relative confusion matrix is as below:

Performance Scores	1	2	3	4
1	4	0	0	0
2	1	3	0	0
3	0	0	69	0
4	0	0	11	1

Table 3. Confusion Matrix